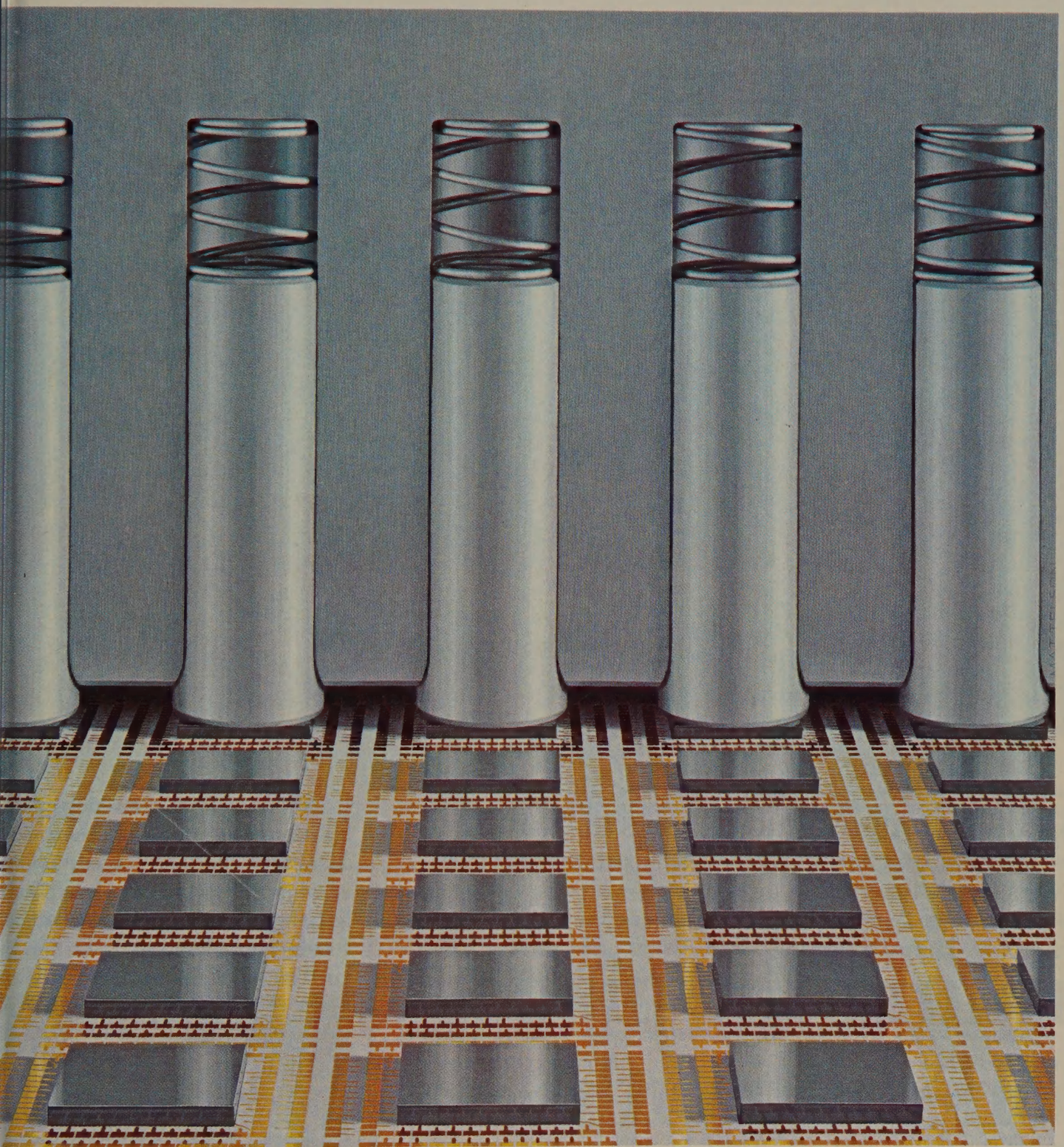
JUL 10 '84

605
S3
V.249
1983
(July-Dec.)

JUL 05 '83

Marygrove College L
Detroit, Michigan 48
PLEASE DO NOT REA

SCIENTIFIC AMERICAN



MICROELECTRONIC PACKAGING

$2.50

July 1983

0 07
74820 08715



An IBM 3081: subatomic physics at Stanford

Finding the pieces in one billionth of a second.

Trying to see how a watch works by slamming two together and examining the pieces is a popular analogy with physicists. It explains how they examine the innermost secrets of subatomic particles.

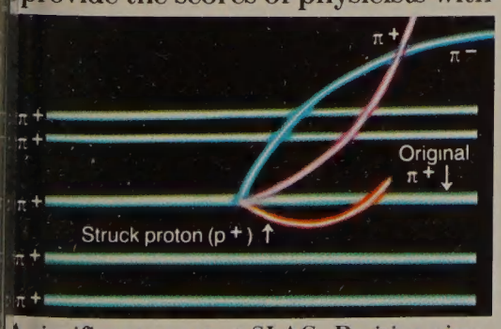
The watches represent matter and antimatter—electrons and positrons that whirl in opposite directions through a 1.3-mile ring at the Stanford Linear Accelerator Center (SLAC). Approaching the speed of light, they collide and generate bursts of subatomic shards that last for only a tiny sliver of time.

An IBM 3081 processor at SLAC has proved a smashing success in finding meaning in millions of bits



of data from these experiments, which may run for weeks or months.

As Charles Dickens, director of SLAC computing services, says, "Our 3081 moves data in and out of its central processor fast enough to provide the scores of physicists with



A significant event at SLAC. Positive pion (π^+) strikes proton (p^+) creating two new mesons, one negative and one positive. Four other pions pass through with no contact.

immediate access to experimental results."

They can think about the problem, and not worry about the computer, with microcode-assisted VM/CMS. With this flexible IBM software and the 3081's reliability, physicists can, with confidence, steal a look for a significant event during the experiment.

Dickens explains, "Before, people might say, 'If I'd only known what was happening, I'd have done things differently.' Now they see the results as they happen."

The 3081 supports 450 terminals for interactive computing under VM/CMS. It performs well in online environments because a new dyadic design allows its two separate processors to work concurrently on a wide range of applications. Its high reliability is due, in part, to an advanced thermal conduction module (TCM).

IBM offers engineering and scientific users extensive support: consultants, education and access to professionals. Tap into our many years of experience.

For an informative brochure, write Dr. Jack W. Hugus, IBM Engineering and Scientific Marketing, 1133 Westchester Avenue, White Plains, NY 10604.

IBM®



"When your traveling companion is worth $2.5 million you better be sure about your transportation."

BOB NORRIS, AMERICAN QUARTER HORSE ASSOCIATION

In the celebrated world of quarter horses, the value of a stallion often runs into the millions. So it's no wonder that Bob Norris, former President of the American Quarter Horse Association, and so many other distinguished figures in equestrian circles, trust the safe transport of their valued cargos to a Jeep® Wagoneer Limited.

There's a special feeling of confidence behind the wheel of a Jeep Wagoneer. There's authority in its power, security in its traction and prestige in its appearance.

With its remarkable Selec-Trac 2-wheel/4-wheel drive system, the Wagoneer artfully masters most any terrain, even under adverse conditions...from summer downpours to winter snowstorms.

Yet, for all its impressive strength and dependability, classic comforts are standard in a Wagoneer Limited. From its plush interior fabrics and leathers to its premium stereo sound system, quality is never compromised.

And because it does so *many* things without compromise, the Jeep Wagoneer Limited makes perfect horse sense, no matter who your traveling companion is.

Jeep® Wagoneer Limited. The Ultimate Wagon.

FROM JEEP CORPORATION

Established 1845 **SCIENTIFIC AMERICAN** July 1983 Volume 249 Number 1

ARTICLES

33 **THE SLOWING OF URBANIZATION IN THE U.S., by Larry Long and Diana DeAre**
The rural population of the country is now growing as fast as the urban, creating "supercities."

42 **HOW CONTINENTS BREAK UP, by Vincent Courtillot and Gregory E. Vink**
Their deformation and rifting, giving rise to a new ocean, is modeled by the Afar depression.

50 **THE SALMONID FISHES AS A NATURAL LIVESTOCK, by Lauren R. Donaldson and Timothy Joyner** Salmon and trout are being "ranched" through knowledge of their biology.

70 **LEARNING IN A MARINE SNAIL, by Daniel L. Alkon**
Experiments with *Hermissenda* trace the nerve circuits that mediate the association of stimuli.

86 **MICROELECTRONIC PACKAGING, by Albert J. Blodgett, Jr.**
The performance of a computer depends partly on how many chips can be got into a small volume.

98 **THE DEVELOPMENT OF PALM LEAVES, by Donald R. Kaplan**
The process is an instructive example of how a complex tissue develops from a simple precursor.

106 **PARTICLES WITH NAKED BEAUTY, by Nariman B. Mistry, Ronald A. Poling and Edward H. Thorndike** A fifth quark, with the "flavor" designated beauty, has been detected.

116 **CREOLE LANGUAGES, by Derek Bickerton**
The similarity of these widely scattered tongues suggests an underlying principle of language.

DEPARTMENTS

6 LETTERS
11 50 AND 100 YEARS AGO
12 THE AUTHORS
14 METAMAGICAL THEMAS
22 BOOKS
60 SCIENCE AND THE CITIZEN
124 THE AMATEUR SCIENTIST
130 BIBLIOGRAPHY

BOARD OF EDITORS Gerard Piel (Publisher), Dennis Flanagan (Editor), Brian P. Hayes (Associate Editor), Philip Morrison (Book Editor), John M. Benditt, Peter G. Brown, Michael Feirtag, Diana Lutz, Jonathan B. Piel, John Purcell, James T. Rogers, Armand Schwab, Jr., Joseph Wisnovsky

ART DEPARTMENT Samuel L. Howard (Art Director), Steven R. Black (Assistant Art Director), Ilil Arbel, Edward Bell

PRODUCTION DEPARTMENT Richard Sasso (Production Manager), Carol Hansen and Leo J. Petruzzi (Assistants to the Production Manager), Carol Eisler (Senior Production Associate), Karen O'Connor (Assistant Production Manager), Carol Albert, Lori Mogol, Martin O. K. Paul, Julio E. Xavier

COPY DEPARTMENT Sally Porter Jenks (Copy Chief), Debra Q. Bennett, Mary Knight, Dorothy R. Patterson

GENERAL MANAGER George S. Conn

ADVERTISING DIRECTOR C. John Kirby

CIRCULATION MANAGER William H. Yokel

SECRETARY Arlene Wright

SCIENTIFIC AMERICAN (ISSN 0036-8733), PUBLISHED MONTHLY BY SCIENTIFIC AMERICAN, INC., 415 MADISON AVENUE, NEW YORK, N.Y. 10017. COPYRIGHT © 1983 BY SCIENTIFIC AMERICAN, INC. ALL RIGHTS RESERVED. PRINTED IN THE U.S.A. NO PART OF THIS ISSUE MAY BE REPRODUCED BY ANY MECHANICAL, PHOTOGRAPHIC OR ELECTRONIC PROCESS, OR IN THE FORM OF A PHONOGRAPHIC RECORDING, NOR MAY IT BE STORED IN A RETRIEVAL SYSTEM, TRANSMITTED OR OTHERWISE COPIED FOR PUBLIC OR PRIVATE USE WITHOUT WRITTEN PERMISSION OF THE PUBLISHER. SECOND-CLASS POSTAGE PAID AT NEW YORK, N.Y., AND AT ADDITIONAL MAILING OFFICES. AUTHORIZED AS SECOND-CLASS MAIL BY THE POST OFFICE DEPARTMENT, OTTAWA, CANADA, AND FOR PAYMENT OF POSTAGE IN CASH. SUBSCRIPTION RATE: $24 PER YEAR IN THE U.S. AND ITS POSSESSIONS; $33 PER YEAR IN ALL OTHER COUNTRIES. POSTMASTER: SEND ADDRESS CHANGES TO SCIENTIFIC AMERICAN, 415 MADISON AVENUE, NEW YORK, N.Y. 10017.

SCIENTIFIC AMERICAN

CORRESPONDENCE

Offprints of more than 1,000 selected articles from earlier issues of this magazine, listed in an annual catalogue, are available at 75 cents each. Correspondence, orders and requests for the catalogue should be addressed to W. H. Freeman and Company, 4419 West 1980 South, Salt Lake City, UT 84121. Offprints adopted for classroom use may be ordered direct or through a college bookstore. Sets of 10 or more Offprints are collated by the publisher and are delivered as sets to bookstores.

Photocopying rights are hereby granted by Scientific American, Inc., to libraries and others registered with the Copyright Clearance Center (CCC) to photocopy articles in this issue of SCIENTIFIC AMERICAN for the flat fee of $2 per copy of each article or any part thereof. Such clearance does not extend to the photocopying of articles for promotion or other commercial purposes. Correspondence and payment should be addressed to Copyright Clearance Center, Inc., 21 Congress Street, Salem, MA 01970. Specify CCC Reference Number ISSN 0036–8733/83. $2.00 + 0.00.

Editorial correspondence should be addressed to The Editors, SCIENTIFIC AMERICAN, 415 Madison Avenue, New York, NY 10017. Manuscripts are submitted at the authors' risk and will not be returned unless they are accompanied by postage.

Advertising correspondence should be addressed to C. John Kirby, Advertising Director, SCIENTIFIC AMERICAN, 415 Madison Avenue, New York, NY 10017.

Subscription correspondence should be addressed to Subscription Manager, SCIENTIFIC AMERICAN, P.O. Box 5969, New York, NY 10017. The date of the last issue on your subscription is shown in the upper right-hand corner of each month's mailing label. For change of address notify us at least four weeks in advance. Please send your old address (if convenient, on a mailing label of a recent issue) as well as the new one.

Name

New Address

Street

City

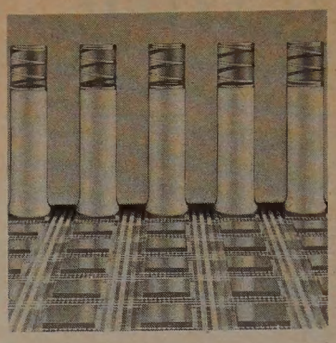
State and ZIP

Old Address

Street

City

State and ZIP



THE COVER

The columns rising above a mosaic floor in the painting on the cover represent part of a cooling module that houses more than 100 silicon chips in the central processing unit of a large digital computer. The chips are mounted face down on a ceramic substrate that supplies some 12,000 terminals for power and signals. The columns, or pistons, are pressed against the back of the chips to carry off heat. In many computers the technology of interconnecting and cooling microelectronic devices is the major factor limiting the performance of the central processor (see "Microelectronic Packaging," by Albert J. Blodgett, Jr., page 86). The module is made by the International Business Machines Corporation.

THE ILLUSTRATIONS

Cover painting by Marvin Mattelson

Page	Source
14–20	William S. Huff, State University of New York at Buffalo
34	National Aeronautics and Space Administration, Toby Carlson, Pennsylvania State University, and Annabelle Carlson
35–41	Albert E. Miller
42	U.S. Geological Survey, EROS Data Center
44	Andrew Tomko
45	Andrew Tomko (*top*); Vincent Courtillot, University of Paris (*bottom*)
46–49	Andrew Tomko
50	Weststock
52	Enid Kotschnig
54–56	Sally Black
57	University of Washington
58	Lauren R. Donaldson, University of Washington College of Ocean and Fishery Sciences
71	Pierre A. Henkart, National Cancer Institute (*top*); Alan M. Kuzirian, Marine Biological Laboratory, Woods Hole, Mass. (*bottom*)
72	Jerome Kuhl
73–74	Daniel L. Alkon, Marine Biological Laboratory, Woods Hole, Mass. (*top*); Jerome Kuhl (*bottom*)
76–84	Jerome Kuhl
87	International Business Machines Corporation
88–90	Gabor Kiss
91	International Business Machines Corporation
92–94	Gabor Kiss
95	International Business Machines Corporation (*top*), Gabor Kiss (*bottom*)
96	International Business Machines Corporation
99	© Walter Chandoha
100	Ilil Arbel
101–102	Donald R. Kaplan, University of California at Berkeley
104	Ilil Arbel
105	Donald R. Kaplan, University of California at Berkeley
107	Allen Beechel
108	Edward Bell
109–115	Allen Beechel
117	Collection of Ray Jerome Baker and Robert E. Van Dyke
118	Ilil Arbel
119–122	Edward Bell
124–128	Michael Goodman

BREAKTHROUGH: ELECTRON BEAM DRILLS HOLES TO GIVE AIRLINES A SPECIAL LIFT.

When air flows over an airplane wing, it creates turbulence and performance-killing drag. We've found a way to reduce drag dramatically and save the airlines millions of dollars in fuel costs.

An electron beam, thinner than a human hair, was used to drill millions of holes in the titanium skin of the wing. (The ordinary paper clip shows how tiny they are.) During flight, a thin layer of air passing over the wing is sucked through the holes to stop turbulence before it develops.

This breakthrough could reduce fuel consumption by as much as 20% on long flights. Testing begins this year. The planes you fly might be using it by the 1990s.

We create breakthroughs not only in aerospace, but in health care, information processing, and energy.

We're McDonnell Douglas.



MCDONNELL DOUGLAS

LETTERS

Sirs:

"Hidden Visual Processes," by Jeremy M. Wolfe [SCIENTIFIC AMERICAN, February], is a provocative challenge to traditional theories of visual perception and a refreshing alternative to commonplace conceptions of human perceptual abilities. The article also raises epistemological issues that deserve further attention. In his effort to show that the human visual system is composed of a number of specialized subsystems acting coordinately to effect our awareness of the world, he argues that "no amount of introspection can make us aware of the subsystems themselves."

Wolfe's experiments are ingenious. His conclusions regarding their cognitive significance are questionable. In his discussion of one typical procedure, that involving "accommodation" (the ability to control the muscles focusing the eye on objects at various distances by changing the shape of the lens) he concludes that since this ability is notably independent of color perception, its discovery is "a finding that no amount of introspection could have suggested about vision." His detailed account of the experimental procedures, however, contains numerous (and unavoidable) references to what the subject saw or thought he saw. For example: "Thus our subjects could readily tell that the isoluminant stimulus was an edge between red and green, but it was impossible for them to bring the edge into focus." How are we to suppose this information was obtained if not by asking the subjects to report their occurrent visual experiences, that is, to "introspect"? These references do not in any way undermine the empirical significance of Wolfe's research, but they do underscore the inconsistency of his methodical repudiation of introspection and his repeated appeal to what are essentially introspective reports of visual experiences *under experimentally directed conditions.* The fact that some perceptual processes may be undetectable under normal or unrestricted conditions is neither evidence nor argument for the assumption that their detection requires a bypass of introspective awareness. On the contrary, the detection of such processes is both psychologically and logically dependent on the ability to produce just such *anomalous* experiences as Wolfe's experiments elicit. Introspection has not been eliminated; it has merely been redirected.

Methodologically, Wolfe's research exemplifies commitments that are common enough, even orthodox in experimental psychology. It is all the more ironic, then, that this discipline, which has been systematically intent on excluding appeal to anything as "subjective" and unquantifiable as introspection, should rely so consistently, if covertly, on the introspective judgments of subjects whose reports are regarded as neutrally objective data when they are produced under the controlled conditions prescribed by the investigators.

ALAN TORMEY

University of Maryland
Baltimore County
Catonsville

Sirs:

I was impressed by the experiments Jeremy M. Wolfe described. I do not, however, agree with his conclusions. At the beginning of his article he is referring to our work on residual vision in brain-injured patients (Ernst C. Pöppel, Richard Held and Douglas Frost in *Nature,* Vol. 243, No. 5405, pages 295–296; June 1, 1973). He writes that we "have found that when people who are perceptually blind because of an injury to the cerebral cortex are asked to direct their eyes toward a spot of light, they do surprisingly well. The subjects report that they cannot see the spot, and so they think they are guessing, but they look in roughly the right direction more often than chance would allow. "Brain-injured people thus show evidence of multiple visual processes: some that are damaged, and hidden ones that remain functional." Although the description of our experimental evidence is correct, I cannot accept Wolfe's conclusion on "multiple visual processes," in the light of his final remark that the visual sense is actually many senses and "that the full range of human visual senses remains to be discovered."

I still believe that there is only *one* visual sense and that what Wolfe is talking about—also in the case where he is referring to our results—are not different visual senses but operations within the one sense that enable the visual system to do the one thing it has been developed for: to recognize objects in the visual world.... A theory of object recognition putting service operations on a lower level is quite different from a theory that defines vision as activity of many visual senses.

ERNST PÖPPEL

Ludwig-Maximilians-Universität
Munich

Sirs:

Alan Tormey raises a common but, I think, incorrect objection to the idea of "hidden" visual processes. He notes that the data are obtained by asking the subject what he or she sees and concludes that since the subject can answer, the processes under study are accessible to conscious visual perception. What the subject "sees" in these experiments and what the experimenter "infers" are two very different things, however. Consider the accommodation experiment. The subject reports on the alignment of two line segments. We, of course, are not interested in the ability to perceive alignment. We use the information to measure the response of the accommodative system, how well the subject was focusing. If we asked directly, "At what distance are you focused?" the subject could not give us a meaningful answer in this experiment.

Our situation is similar to that of the physician. If you have an abdominal pain, the doctor does not ask you which of your various internal organs is causing the problem. Instead you are asked where it hurts, and from this information the doctor can try to diagnose the underlying cause. The apparent locus of the pain may be some distance from the site of the problem. No matter; the doctor is not interested in your experience of pain but only in what it can reveal about the state of your health.

In the experiments described in my article we used two different line-alignment tasks and a measure of apparent self-motion. From this we inferred that the accommodative system is colorblind and that there are at least three different binocular processes in human vision. The specific tasks we used were irrelevant to our conclusions. The conclusions cannot be derived from conscious experience either in or out of the experiment. This is so even though the conclusions are based on interpretations of what our subjects report seeing.

Ernst Pöppel is, I think, correct when he makes object recognition the "primary task" of the visual system. As he says, under normal circumstances functions such as accommodation serve to make the task easier. Nevertheless, the subservient process may still be treated as separate, self-contained and "hidden" when it is studied in the laboratory. The accommodative process is responsive to certain stimuli and gives rise to certain responses. In the laboratory it can be studied in the way we would study the vision of an animal. We present stimuli and make inferences on the basis of a motor output. I would not question Dr. Pöppel's main point, that there is a unified visual sense. I would suggest that the unity of vision may be like the unity of an army. At one level it moves with a single purpose. On more detailed examination it is composed of a number of autonomous parts acting together.

JEREMY M. WOLFE

Department of Psychology
Massachusetts Institute of Technology
Cambridge

Show her she's the reason it's never been lonely at the top.



Your jeweler can show you many extraordinary diamond solitaires 1 carat or more. The ring shown features 1.67 carat brilliant cut diamond.



A carat or more—one in a million.

Every diamond is rare.
But of all diamonds found, a solitaire of a carat or more is only one in a million.

And, like love, becomes more precious with time.

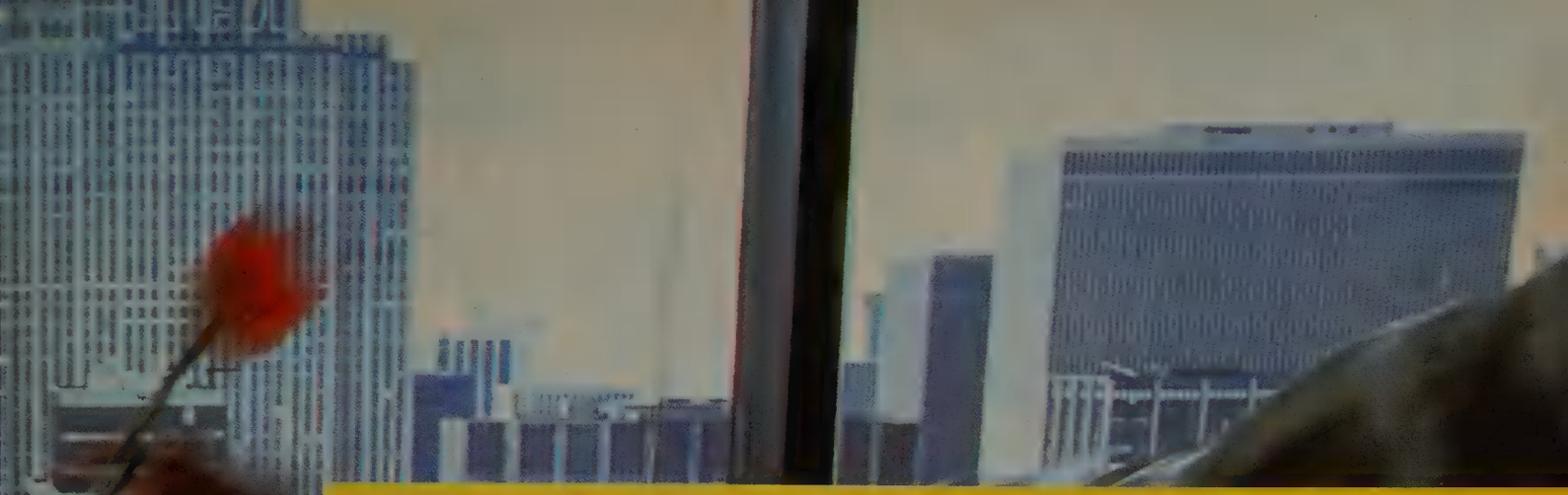
A miracle among miracles.
Born from the earth. Reborn on a woman.

The extraordinary diamond of a carat or more.

Show the world you couldn't have made it without her.

A diamond is forever. De Beers


VICTOR



Victor. How a first name can make it big in the business world.

What does it take these days to be successful?

First, speak several languages. Have a large and infallible memory. Adapt at lightning speed. Stay accessible 24 hours a day. Be practical. And, above all, remain simple and inexpensive.

That's exactly why Victor® is a leader. The leading 16-bit microcomputer in Europe.

Start with the fact that Victor speaks virtually all Western languages. Fluently. And, you can adapt Victor's keyboards to either language or business specific applications.

Then, Victor has a large and useful memory. That's vital.

And Victor is highly flexible and adaptable. With appropriate software, Victor can perform thousands of jobs at electronic speed. In fact, Victor's extensive software library allows you to put Victor to work in your office, department or plant immediately. Productively.

Finally, Victor is cost-conscious. When it comes to price and performance, you can't buy more computer for less.

A success in the business world? Absolutely. To find out how a first name can make it big in your business, contact Victor at the address or phone number below.

VICTOR®

The computer you call by name.

VICTOR TECHNOLOGIES INC. 380 EL PUEBLO ROAD. SCOTTS VALLEY. CA 95066 (408) 438-6680.
VICTOR TECHNOLOGIES EUROPE. NINOOFSESTEENWEG 71. 1750 SCHEPDAAL. BELGIQUE. (2)569.5500.

We don't care which computer you buy.

We'll help you get the most out of it.



CompuServe puts a world of information, communications, and entertainment at your fingertips.

CompuServe is the easy to use videotex service designed for the personal computer user and managed by the communications professionals who provide business information services to over one fourth of the FORTUNE 500 companies.

Subscribers get a wealth of useful, profitable, or just plain interesting information like national news wires, electronic banking and shop at home services, and sophisticated financial data. Plus, a communications network for electronic mail, a bulletin board for selling, swapping, and personal notices and a multi-channel CB simulator.

You get games on CompuServe, too. Classic puzzlers, educational, sports and adventure games and fantastic space games featuring MegaWars, the "ultimate computer conflict."

To learn more about CompuServe, call toll-free, 800-848-8199, for an illustrated guide to the CompuServe Information Service. The videotex service for you, no matter which computer you buy.

CompuServe

Consumer Information Service
2180 Wilson Rd., Columbus, Ohio 43228

800-848-8199 In Ohio Call 614-457-8650

An H&R Block Company

50 AND 100 YEARS AGO

SCIENTIFIC AMERICAN

JULY, 1933: "Professor R. A. Millikan's suggestive theory of the origin of cosmic rays is based on the assumption that the rays are electromagnetic radiations, or photons, similar in type to X rays and gamma rays but of shorter wavelength. About five years ago two German physicists, Walter Bothe and Werner Kolhörster, did an experiment with counting tubes that convinced them the cosmic rays are electrically charged particles. If this conclusion is correct, it means there should be a difference in the intensity of the rays over different parts of the globe. The earth acts as a huge magnet, which should deflect the electrified particles as they shoot toward it. The effect should be least near the magnetic poles and greatest near the Equator, resulting in an increasing intensity from the Equator toward the poles. A group led by Arthur H. Compton of the University of Chicago has organized nine different expeditions over the past 18 months, going to different parts of the globe to measure cosmic rays. When the results of these expeditions are brought together, it is found that the cosmic-ray intensity near the poles is about 15 per cent greater than that near the Equator. These results show that a considerable part, at least, of the cosmic rays consists of electrified particles."

"Invisible 'black light' was described by Samuel G. Hibben of the Westinghouse Electric and Manufacturing Company at a recent meeting of the Illuminating Engineering Society in New York City. 'The "black light,"' said Mr. Hibben, 'is ultra-violet radiation, and it is 99 per cent free of visible light. It is produced by two new black-bulb lamps, one consuming two amperes, the other five. They are made of a special cobalt glass. The bulbs of these lamps absorb 99 per cent of the visible light and transmit 80 to 85 per cent of the ultra-violet wanted. The radiations are relatively long in wavelength, in the range of 3,200 to 4,000 angstrom units. This long-wave ultra-violet lends itself to many photographic and fluorescent effects.'"

"From the time the *Graf Zeppelin* was first placed in service in 1928 until the end of last year the airship made 290 voyages, including 33 ocean crossings, traveling 330,000 miles in 5,370 hours, at an average speed of 61 miles per hour, and carrying almost 17,000 passengers, 35,000 pounds of mail and 80,000 pounds of freight, without accident or mishap, a remarkable demonstration of safety and dependability. A number of improvements are incorporated in a new leviathan of the air that is now nearing completion in Germany and that is to supplement the South American service of the *Graf Zeppelin*. The airship and the airplane differ in that the airplane is primarily a fast, comparatively short-radius craft, whereas the airship is slower and reaches full efficiency only on long voyages, particularly across oceans. In the aviation of the future there is every reason to believe that the airplane will find its place as the carrier for relatively short distances—even making contact with the dirigible, as has already been successfully accomplished during the regular Germany-Brazil service of the *Graf Zeppelin*—and that the dirigible will be the vehicle of transcontinental and intercontinental travel."

"Former industrial workers, storekeepers, barbers, clerks, electricians, miners—these and other unemployed men from many other walks of life are the direct beneficiaries of the new Federal reforestation program. A quarter of a million men will be paid for their reforestation work, will be assured their subsistence for the next few months and will get the healthy, outdoor work that will help to make them physically fit and restore their mental equilibrium weakened by privation and despair. No less direct will be the benefit to the entire American people."

SCIENTIFIC AMERICAN

JULY, 1883: "Nearly all the petroleum that goes into the world's commerce is produced in a district about 150 miles long, with a varying breadth of from one mile to 20 miles, lying mainly in the state of Pennsylvania but lapping over a little on its northern edge into the state of New York. This region yielded, in 1882, 31,398,750 barrels. A little petroleum is obtained in West Virginia, a little at isolated points in Ohio and a little in the Canadian province of Ontario. There is also a small field in Germany, a larger one, scantily developed, in southern Russia and one still larger, perhaps, in India. The total production of all the fields, outside of the region here described, is but a small fraction in the general account. Furthermore, the oil of these minor fields, whether in America or the Old World, is of inferior quality."

"The Westinghouse brake is fitted upon the engine, tender and each vehicle of the train. Air compressed by a pump on the locomotive to, say, 70 lb. or 80 lb. to the square inch fills the main reservoir on the engine, and flowing through the driver's brake valve and main pipe it also charges the supplementary reservoirs throughout the train. When the train is running, uniform air pressure exists throughout its length, that is to say, the main reservoir on the engine, the pipe from one end of the train to the other, the valves and supplementary reservoirs on each vehicle are all charged ready for work, the brake cylinders being empty and the brakes off. The essential principle of the system is that maintaining the pressure keeps the brakes off, but letting the air escape from the brake pipe, on purpose or by accident, instantly applies them. It follows, therefore, that the brake may be applied by the driver or any of the guards, or if necessary by a passenger, by the separation of a coupling, or the failure of or injury to a vital part of the apparatus; since the brake on each vehicle is complete in itself and independent, should the apparatus on any one car be torn off, the brakes will nevertheless remain applied for almost any length of time in the rest of the train."

"It is an old and probably a true saying that every element could be detected everywhere had we sufficiently delicate tests for it. Early observations had prepared William Crookes, F.R.S., for the wide distribution of yttrium, and no sooner had Mr. Crookes developed his method of radiant-matter spectroscopy than the exquisite sensitiveness of this spectrum test forced itself on his notice when he sought for yttrium in various minerals. Pink coral contains one part of yttrium in 200 parts; strontianite, one part of yttrium in 500 parts; chondrodite from Mount Somma, one part in 4,000; calcite, one part in 10,000; ox bone, one part in 10,000; an earthy meteorite (Alfianello), one part in 100,000, and tobacco ash, one part in 1,000,000."

"Should women ride like men? Although it may not appear to be the case, the seat a woman enjoys on a side-saddle is fully as secure, and not nearly as irksome, as that which a man has to maintain, unless he simply balances himself and does not gripe the sides of his horse with either the knee or the side of the leg. It is curious to note the different ways in which the legs of men who pass much time in the saddle are affected. Riding with a straight leg and a long stirrup almost invariably produces what are popularly called knocked knees. On the other hand, riding with a short stirrup produces bowed legs. No deformity necessarily follows the use of the side-saddle if the precaution is taken with growing girls to change sides on alternate days, riding on the left side one day and on the right the next."

80 PROOF · CALVERT DISTILLERS CO., NYC

If your hot stock tip just turned cold, buy into a rewarding change of pace.



John Jameson
Imported Irish Whiskey

THE AUTHORS

LARRY LONG and DIANA DEARE ("The Slowing of Urbanization in the U.S.") are members of the staff of the U.S. Bureau of the Census. They are utilizing the results of the census of 1980, among other data, to study the distribution of population and economic activity in the U.S. Long is at the bureau's Center for Demographic Studies. He received his undergraduate education at the University of Texas at Austin, going on to obtain his Ph.D. in sociology there in 1969. He did postdoctoral work at the University of Pennsylvania before moving to the Census Bureau while the results of the census of 1970 were being tabulated. DeAre is with the bureau's Population Division. She got her B.A. in 1971 at DePaul University. Her M.A. (1973) and Ph.D. in geography (1976) are from the University of Texas at Austin. She joined the bureau in 1975.

VINCENT COURTILLOT and GREGORY E. VINK ("How Continents Break Up") are geophysicists who did independent work on the subject of their article before collaborating on writing it. Courtillot is professor of geophysics and chairman of the department of earth sciences at the University of Paris. He attended the Paris School of Mines as an undergraduate. He came to the U.S. to pursue his studies, earning his M.S. in geophysics at Stanford University in 1972 before returning to France; his Ph.D. in geophysics was awarded in 1977 by the University of Paris. Courtillot was a member of the recent French-Chinese geological expedition to Tibet. Vink is a graduate student at Princeton University. He received his B.A. in 1979 from Colgate University. He is writing a dissertation on the plate-tectonic evolution of the Norwegian-Greenland Sea and the Arctic Ocean.

LAUREN R. DONALDSON and TIMOTHY JOYNER ("The Salmonid Fishes as a Natural Livestock") are experts in fisheries. Donaldson is professor emeritus of fisheries at the University of Washington. He got his B.S. from Intermountain Union College. He obtained his M.S. and Ph.D. from the University of Washington before joining the faculty of the College of Ocean and Fishery Sciences there in 1932. He has made a significant contribution to the large fisheries program at Washington. He developed a brood stock of rainbow trout with rates of growth, survival and egg production more than 20 times those of previous trout stock. He directed the work that established runs of salmon and steelhead trout to the university's holding ponds. Donaldson served for many years as director of the Laboratory of Radiation Biology at Washington, which evaluates the biological effects of radioactivity. Joyner earned his Ph.D. in fisheries in 1962, partly for work done at the Laboratory of Radiation Biology. He went to Brown University as an undergraduate before serving in the Marine Corps in the Pacific in World War II. After getting his doctorate he went to work for the National Marine Fisheries Service (then the Bureau of Commercial Fisheries). He retired in 1975 and is now a writer and consultant on fisheries. Joyner writes that he is "following with great interest a project I helped to start while I was with the National Marine Fisheries Service: rearing selected stocks of Atlantic salmon in saltwater pens in Puget Sound to help provide eggs for the restocking of rivers in New England."

DANIEL L. ALKON ("Learning in a Marine Snail") is research medical officer and chief of the section on neural systems at the National Institute of Neurologic and Communicative Disorders and Stroke. He is also on the staff of the Marine Biological Laboratory in Woods Hole, Mass.; his laboratory is at Woods Hole. His B.S. (1965) is from the University of Pennsylvania. His M.D. (1969) is from the Cornell University Medical College. After completing his internship he went to the National Institutes of Health. Alkon's research interests include neuropharmacology, sensory physiology and animal behavior.

ALBERT J. BLODGETT, JR. ("Microelectronic Packaging"), is manager of the manufacturing research laboratory at the Thomas J. Watson Research Center of the International Business Machines Corporation. He was graduated from Yale University with a B.E. in electrical engineering in 1960. In the same year he moved to the product-development laboratory at IBM. Two years later he left for Stanford University on an IBM resident fellowship; at Stanford he earned his M.S. and Ph.D. degrees in electrical engineering before returning to IBM in 1966. Among other projects, Blodgett was responsible for the design of the solid-logic-technology modules that are the basic logic devices in the S/360 data processor and the development of the multilayered ceramic modules utilized in the 43XX and 308X data-processor units.

DONALD R. KAPLAN ("The Development of Palm Leaves") is professor of botany at the University of California at Berkeley. He got his A.B. in

1960 at Northwestern University. His Ph.D. in botany, granted in 1965, is from Berkeley. In 1965 he served as a National Science Foundation fellow at the Royal Botanical Garden at Kew in England. He returned to the U.S. to take up a job in the department of organismic biology at the University of California at Irvine. He went to Berkeley in 1968. In addition to Kaplan's botanical interests he is a free-lance writer on (and photographer of) railroad trains.

NARIMAN B. MISTRY, RONALD A. POLING and EDWARD H. THORNDIKE ("Particles with Naked Beauty") are members of a group of 75 physicists from seven universities who are investigating the physics of beauty-flavored mesons with the detector named CLEO at the Cornell Electron Storage Ring (CESR). Mistry is senior physicist at the Newman Laboratory of Nuclear Studies at Cornell University. A native of India, he received his undergraduate education at St. Xavier's College in Bombay. He came to the U.S. in 1958 to continue his training in physics; Columbia University awarded him a Ph.D. in 1963. He has been at Cornell since then. Poling is research associate in physics and astronomy at the University of Rochester. His B.S. (1976) is from the State University of New York at Buffalo and his Ph.D. (1981) is from Rochester. Thorndike is professor of physics at Rochester. He was graduated from Wesleyan University in 1956 with an A.B. and went on to earn his M.S. at Stanford University in 1957. His Ph.D. in physics (1960) is from Harvard University. He has served as spokesman of the CLEO collaboration since 1981.

DEREK BICKERTON ("Creole Languages") writes: "I acquired my B.A. from the University of Cambridge in 1949 but did not start my academic career until 1964, when I taught English literature at the University College of Cape Coast in Ghana. In 1966 and 1967 I studied linguistics at the University of Leeds before becoming senior lecturer in the English language at the University of Guyana, where I remained for four years. I taught briefly at the University of Lancaster in England and joined the University of Hawaii at Manoa in the fall of 1972. In the British system it isn't necessary to have a doctorate but in the American it is, and so I took advantage of a regulation that enables graduates of Cambridge to submit published work instead of a regular dissertation. The work I submitted was *Dynamics of a Creole System* (Cambridge University Press, 1975). I obtained my doctorate in 1976. I have held the rank of full professor at Hawaii since the same year. I have also written several novels, the most recent of which was *King of the Sea* (Random House, 1980)."

WHAT'S BETTER THAN SPEED READING?

SPEED LEARNING

(SPEED PLUS COMPREHENSION)

Speed Learning is replacing speed reading. It's easy to learn . . . lasts a lifetime . . . applies to everything you read . . . and is the only accredited course with special editions for professional reading.

Do you have too much to read and too little time to read it? Do you mentally pronounce each word as you read? Do you frequently have to go back and re-read words or whole paragraphs you just finished reading? Do you have trouble concentrating? Do you quickly forget most of what you read?

If you answer "yes" to any of these questions—then here at last is the practical help you've been waiting for. Whether you read for business or pleasure, school or college, you will build exceptional skills from this major breakthrough in effective reading, created by Dr. Russell Stauffer at the University of Delaware.

Not just "speed reading"—but speed reading-thinking-understanding-remembering-and-learning

The new *Speed Learning Program* shows you step-by-proven-step how to increase your reading skill and speed, so you understand more, remember more and use more of everything you read.

Imagine the new freedom you'll have when you learn how to dash through all types of reading material *at least* twice as fast as you do now, and with greater comprehension. Think of being able to get on top of the avalanche of newspapers, magazines and correspondence you have to read . . . finishing a stimulating book and retaining facts and details more clearly and with greater accuracy than ever before.

What makes Speed Learning so successful?

In just a few spare minutes a day of easy reading and exciting listening, you discover an entirely new way to read and think—a radical departure from anything you have ever seen or heard about. Research shows that reading is 95% *thinking* and only 5% eye movement. Yet most of today's speed reading programs spend their time teaching you rapid eye movement (5% of the problem) and ignore the most important part (95%) *thinking*. In brief, *Speed Learning* gives you what speed reading *can't*.

Executives, students, professional people, men and women in all walks of life from 15 to 70 have benefited from this program. Speed Learning is a fully accredited course . . . costing only ⅕ the price of less effective speed reading classroom courses. Now you can examine the same easy, practical and proven methods at home . . . in spare time . . . without risking a penny.

Examine Speed Learning FREE for 15 days

You will be thrilled at how quickly this program will begin to develop new thinking and reading skills. After listening to just one cassette and reading the preface you will quickly see how you can achieve increases in both the speed at which you read and in the amount you understand and remember.

You must be delighted with what you see or you pay nothing. Examine this remarkable program for 15 days. If, at the end of that time you are not convinced that you would like to master Speed Learning, simply return the program and owe nothing.



learn INCORPORATED Dept. DZ-22, 113 Gaither Drive, Mount Laurel, NJ 08054

YES! Please send the materials checked below:

☐ Standard Speed Learning program @ $99.95 plus $5 postage and handling

☐ Special Edition Speed Learning program @ $125 plus $5 postage and handling (check one below):

☐ Medical ☐ Data Processing ☐ Science/Engineering ☐ Management ☐ Finance/Accounting

Check the method of payment below: NJ Residents add 6% sales tax

☐ Check or money order enclosed

☐ Charge my credit card:

☐ Visa ☐ Master Card ☐ American Express

Card #_____ Exp. Date_____

I understand that I may examine the materials for 15 days. At the end of that time, if I am not delighted in every way I may return them for a full refund. No questions asked.

Name_____

Address_____

City_____ State_____ Zip_____

Signature_____

If you don't already own a cassette player, you may order this Deluxe Cassette Recorder for only $49.95. (Includes handling and delivery.)

Check here to order ☐



Outside USA add $10 per item—Airmail extra

METAMAGICAL THEMAS

Parquet deformations: patterns of tiles that shift gradually in one dimension

by Douglas R. Hofstadter

What is the difference between music and visual art? If someone asked me this question, I would have no hesitation in responding. To me the major difference is temporality. Works of music intrinsically involve time; works of visual art do not. More precisely, pieces of music consist of sounds intended to be played and heard in a specific order and at a specific speed. Music is therefore fundamentally one-dimensional; it is tied to the rhythms of our existence. Works of visual art, in contrast, are generally two- or three-dimensional. Paintings and sculpture seldom have any intrinsic "scanning order" built into them that the eye must follow. Mobiles and other pieces of kinetic art may change over time, but often without any specific initial state or final state or intermediate states. You are free to come and go as you please.

There are, of course, exceptions to this generalization. European art has grand friezes and historic cycloramas, and Oriental art has intricate pastoral scrolls up to hundreds of feet long. These types of visual art impose a temporal order and speed on the scanning eye. There is a starting point and a final point. Usually, as in stories, these points represent states of relative calm, particularly at the end. In between, various types of tension are built up and resolved in an idiosyncratic but pleasing visual rhythm. The calmer end states are usually orderly and visually simple, whereas the tenser intermediate states are usually more chaotic and visually confusing. If you replace "visual" by "aural," virtually the same can be said of music.

I have been fascinated for many years by the idea of trying to capture the essence of the musical experience in visual form. I have my own ideas about how this can be done; in fact, I spent several years working out a form of visual music. By no means, however, do I think there is a unique or best way to carry out this task of "translation," and indeed I have often wondered how others might attempt to do it. I have seen a few such attempts, but most of them struck me as being unsuccessful. One striking counterexample is the set of "parquet deformations" meta-composed by William S. Huff, professor of architectural design at the State University of New York at Buffalo.

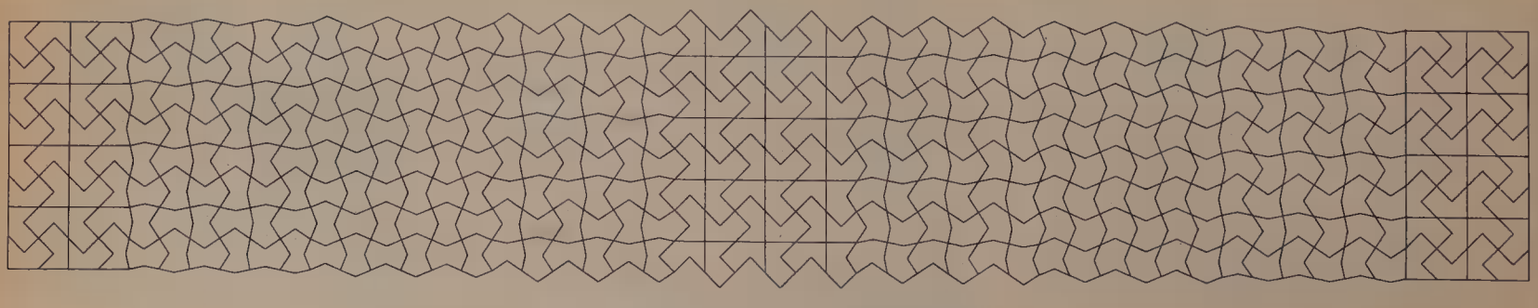
I say "meta-composed" for good reason. Huff himself has never executed a single parquet deformation. He has elicited hundreds of them, however, from his students, and in so doing he has brought this form of art to a high degree of refinement. He might be likened to the conductor of a fine orchestra. Although the conductor makes no sound in the course of a performance, we give much credit to the person doing the job for the quality of the sound. We can only guess at how much preparation and coaching went into the performance.

So it is with William Huff. For 20 years his students in Buffalo and also at Carnegie-Mellon University have been prodded into flights of artistic inspiration, and it is thanks to Huff's vision of what constitutes quality that some beautiful results have emerged. Not only has he elicited outstanding work from students; he has also carefully selected what he thinks are the best pieces. These he is keeping in archives. For these reasons I shall be referring to "Huff's creations" but always in this more indirect sense of meta-creations.

Not to take credit from the students who executed the individual pieces, there is a larger sense of the term "credit" that goes exclusively to Huff, the person who has shaped this entire art form himself. Let me use an analogy. Gazelles are marvelous beasts, yet it is not they themselves but the selective pressures of evolution that are responsible for their species' unique and wondrous qualities. Huff's judgments and comments have here played the role of those impersonal evolutionary selective pressures, and out of them has been molded a living and dynamic tradition, a "species" of art exemplified and extended by each new instance.

All that remains to be said by way of introduction is the meaning of "parquet deformation." Actually it is nearly self-explanatory. Traditionally a parquet is a regular mosaic made out of inlaid wood on the floor of an elegant room, and a deformation is a deformation. Huff's parquets are more abstract: they are regular tessellations (or tilings) of the plane, ideally drawn with zero-thickness line segments and curves. The deformations are not arbitrary but must satisfy two basic requirements: (1) there must be change only in one dimension, so that it is possible to see a temporal progression in which one tessellation gradually becomes another, and (2) at each stage the pattern must constitute a regular tessellation of the plane, that is, there must be a unit cell that could combine with itself so that it could cover an infinite plane exactly.

From this very simple idea emerge some stunningly beautiful creations. Huff explains that he was originally inspired, back in 1960, by the M. C. Escher woodcut "Day and Night." In that work forms of birds tiling the plane are gradually distorted (as the eye scans downward) until they become diamond-shaped, looking like the checkerboard pattern of cultivated fields seen from the air. Escher of course became famous for



A parquet deformation titled "Fylfot Flipflop"

his tessellations, both pure and distorted, as well as for the other haunting games he played with art and reality.

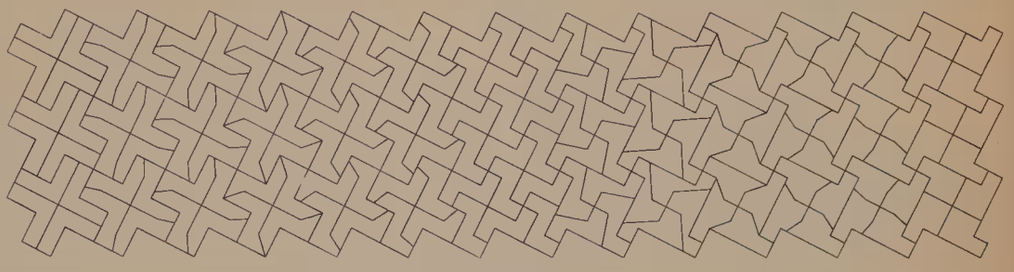
Whereas Escher's tessellations are almost always based on animal forms, Huff decided to limit his scope to purely geometric forms. In a way that is like a decision by a composer to follow austere musical patterns and to totally eschew anything that might conjure up a "program" (that is, some kind of image or story behind the sounds). An effect of this decision is that the beauty and visual interest must come entirely from the complexity and subtlety of the interplay of abstract forms. There is nothing to "charm" the eye, as there is with pictures of animals. There is only the unembellished perceptual experience.

Because of the linearity of this form of art, Huff has likened it to visual music. He writes: "Although I am spectacularly ignorant of music, tone-deaf and hated those piano lessons (yet can be enthralled by Bach, Vivaldi or Debussy), I have the students 'read' their designs as I suppose a musician might scan a work: the themes, the events, the intervals, the number of steps from one event to another, the rhythms, the repetitions (which can be destructive, if not totally controlled, as well as reinforcing). These are principally temporal, not spatial, compositions (although all predominantly temporal compositions have, of necessity, an element of the spatial and vice versa—e.g., the single-frame picture is the basic element of the moving picture)."

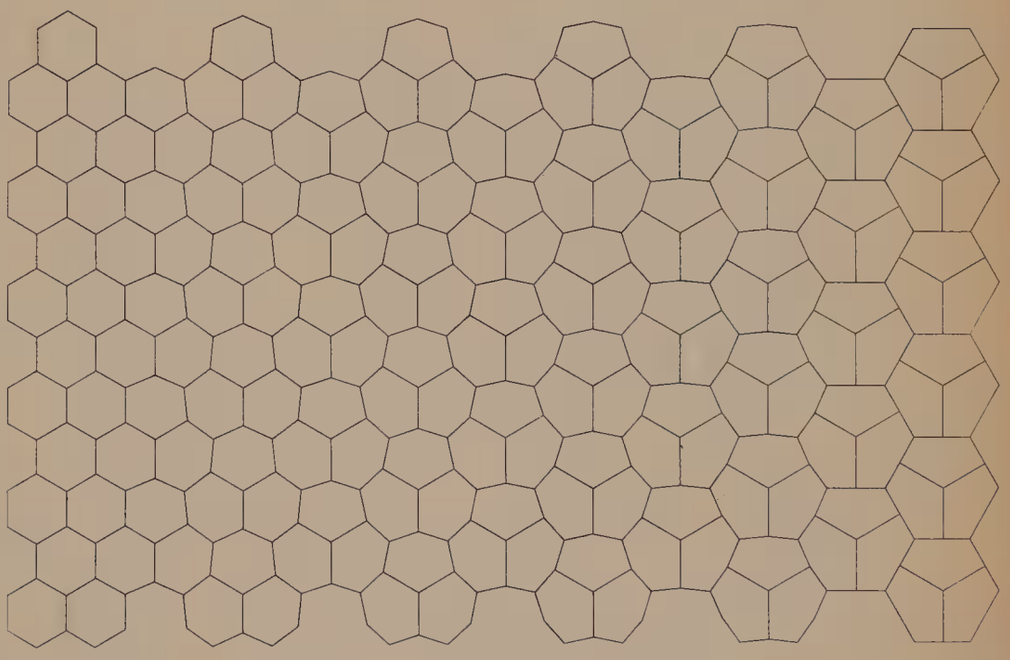
What are the basic elements of a parquet deformation? First there is the class of allowed parquets. On this Huff writes: "We play a different (or rather, tighter) game from Escher's. We work with only *A* tiles (i.e., congruent tiles of the same handedness). We do not use, as he does, *A* and *A'* tiles (i.e., congruent tiles of both handednesses), although an exception to this rule is the example called Dual. Finally, we don't use *A* and *B* tiles (i.e., two different interlocking tiles), since two such tiles can always be seen as subdivisions of a single larger tile."

The other basic element is the repertory of standard deforming devices. Typical devices include lengthening or shortening a line; rotating a line; introducing a "hinge" somewhere inside a line segment so that it can "flex"; introducing a "bump" or "pimple" or "tooth" (a small intrusion or extrusion having a simple shape) in the middle of a line or at a vertex; shifting, rotating, expanding or contracting a group of lines that form a natural subunit, and variations on these themes. To understand these descriptions you must realize that a reference to "a line" or "a vertex" is actually a reference to a line or a vertex inside a unit cell, and therefore when *one* such line or vertex is altered, all the corresponding lines or vertexes that play the same role in the copies of that cell undergo the same change. Since some of those copies may be at 90 degrees (or other angles) with respect to the master cell, one locally innocent-looking change may induce changes at corresponding spots, resulting in unexpected interactions whose visual consequences can be quite exciting.

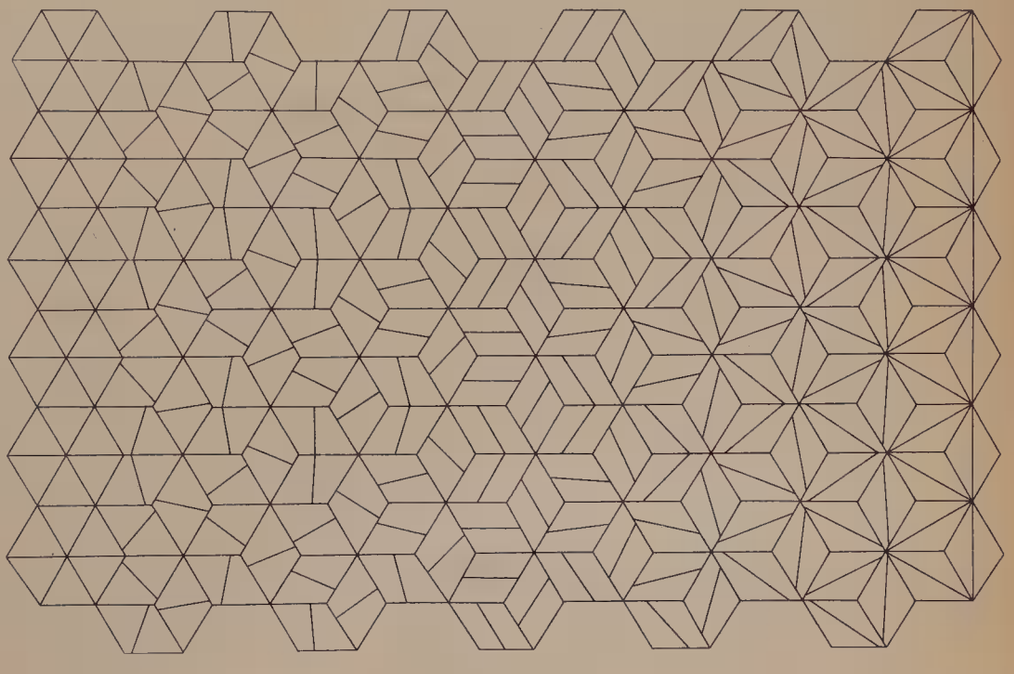
Without further ado, let us proceed to examine some specific pieces. Look at the one titled "Fylfot Flipflop" [*opposite page*]. It is an early one, exe-



"Crossover"

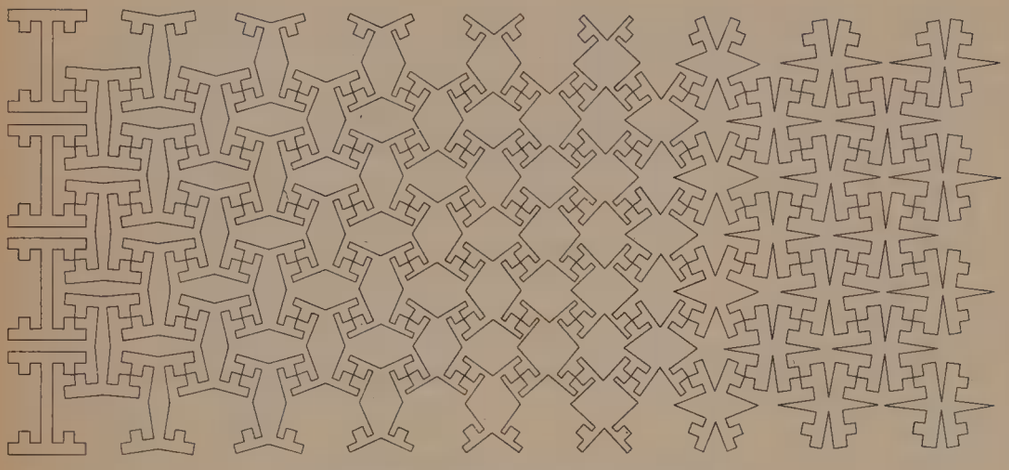


"Dizzy Bee"

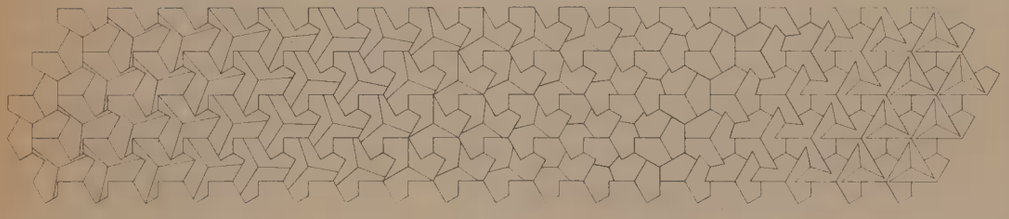


"Consternation"

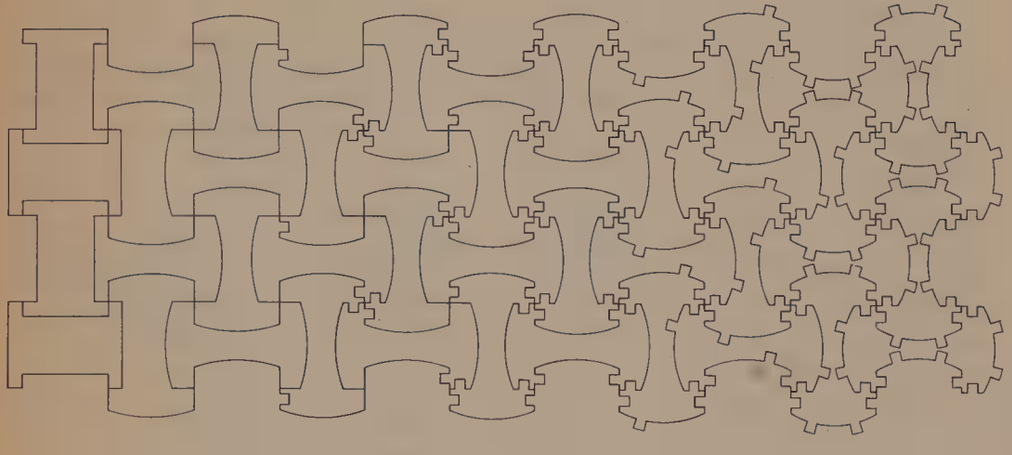
cuted by Fred Watts at Carnegie-Mellon in 1963. If you simply let your eye skim across the top line, you will get the distinct sensation of scanning a tiny mountain range. At each edge you begin with a perfectly flat plain and then move into gently rolling hills, which become taller and steeper, eventually turning into jagged peaks; past the center point these start to soften into lower foothills, which gradually tail off into the plain again. This much is obvious even at a casual glance. Subtler to see is the line just below, whose zigging and zagging is 180 degrees out of phase with the top line. Notice that in the very center that line is completely at rest: a perfectly horizontal stretch flanked on each side by increasingly toothy regions. Below it there are seven more horizontal lines. Thus if one completely filtered out the vertical lines, one would see nine horizontal lines stacked above one another, the odd-numbered ones jagged in the center, the even-numbered ones smooth in the center.



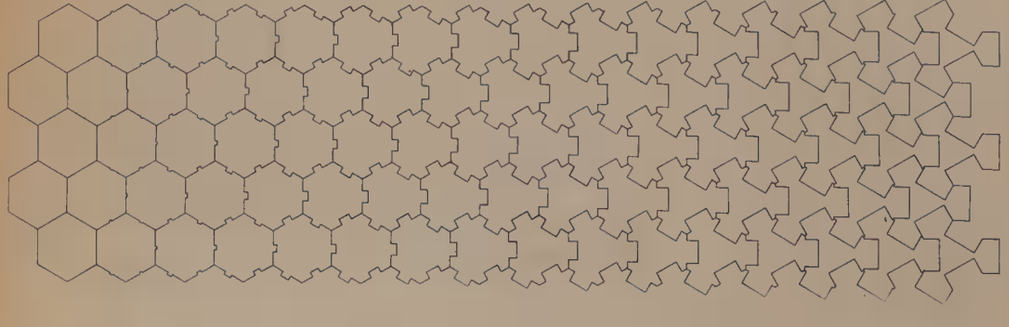
"Oddity out of Old Oriental Ornament"



"Y Knot"



"Crazy Cogs"



"Trifoliolate"

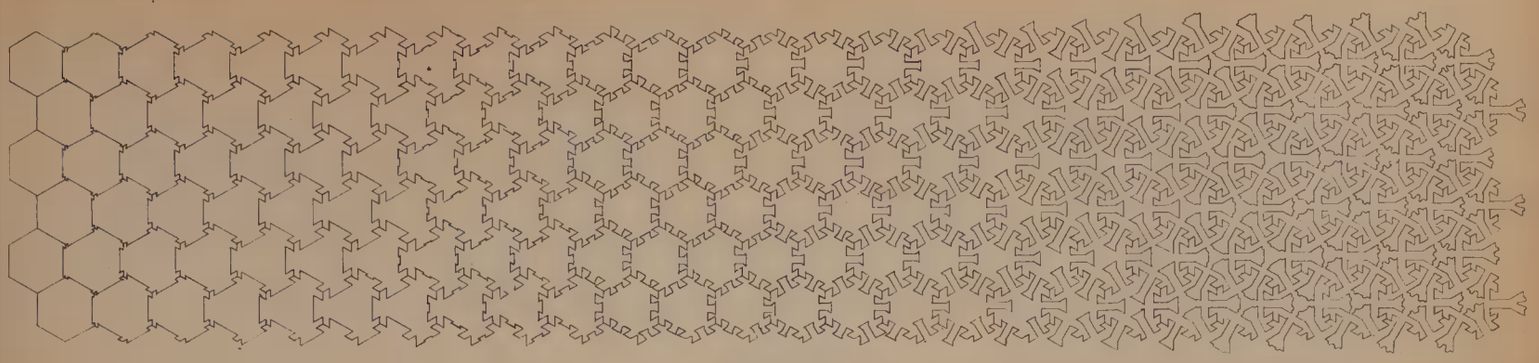
What about the vertical lines? Both the left-hand and the right-hand borderlines are perfectly straight vertical lines. Their immediate neighbors, however, are as jagged as possible, consisting of repeated 90-degree bends, back and forth. The next vertical line nearer the center is practically straight up and down again. Then there is a wavy line again, and so on. As you move across the picture you see that the jagged lines gradually get less jagged and the straight ones get increasingly jagged, until in the middle the roles are completely reversed. The process then continues, so that by the time you have reached the other side the lines are back to normal. If you could filter out the horizontal lines, you would see a simple pattern of quite jagged lines alternating with less jagged ones.

When these two extremely simple, independent patterns—the horizontal and the vertical—are superposed, what emerges is an unexpectedly rich perceptual feast. At the far left and right the eye picks out fylfots—that is, swastikas—of either handedness contained inside perfect squares. In the center the eye immediately sees that the central fylfots are all gone, replaced by perfect crosses inside pinwheels.

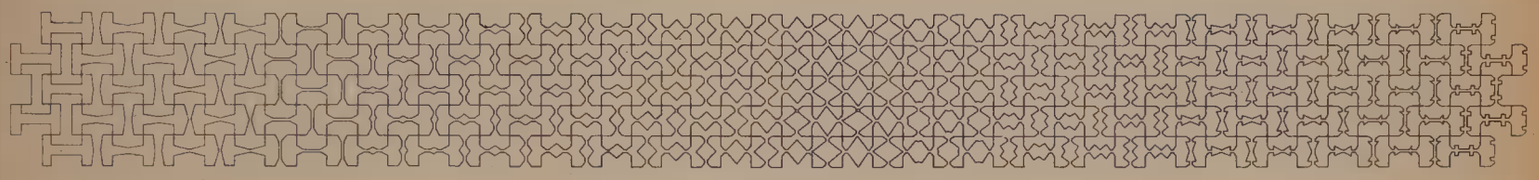
And then a queer perceptual reversal takes place. If you just shift your focus of attention diagonally by half a pinwheel, you will notice that there is a fylfot right there before your eyes! In fact, suddenly fylfots appear all over the central section where before you had been seeing only crosses inside pinwheels. And conversely, of course, now when you look at either end, you will see pinwheels everywhere with crosses inside them. No fylfots! It is an astonishingly simple design, yet the effect catches nearly everyone off guard.

This is a simple example of the ubiquitous visual phenomenon called "regrouping," in which the boundary line of the unit cell shifts so that structures jump out at the eye that before were completely submerged and invisible, whereas conversely, of course, structures that a moment ago were completely obvious are now invisible, having been split into separate conceptual pieces by the act of regrouping, or shift of perceptual boundaries. It is both a perceptual and a conceptual phenomenon, a delight to the subtle combination of eye and mind that is most sensitive to pattern.

For another example of regrouping take a look at "Crossover," executed



"Arabesque"



"Razor Blades"

by Richard Lane, also at Carnegie-Mellon in 1963 [*page 15*]. Something really amazing happens in the middle, but I won't tell you what. Just find it yourself by careful looking.

By the way, there are still features left to be explained in "Fylfot Flipflop." At first it appears to be mirror-symmetrical. For instance, all the fylfots at the left end are spinning counterclockwise, while all the ones at the right end are spinning clockwise. So far, so symmetrical. In the middle, however, all the fylfots go counterclockwise. This surely violates the symmetry. Furthermore, the one-quarter-way and three-quarter-way stages of the deformation, which ought to be mirror images of each other, bear no resemblance at all to each other. Can you figure out the logic behind this subtle asymmetry between the left and right sides?

This piece also illustrates yet another way parquet deformations resemble music. A unit cell—or rather, a vertical cross section consisting of a stack of unit cells—is analogous to a measure in music. The regular pulse of a piece of music is given by the repetition of unit cells across the page. And the flow of a melodic line across measure boundaries is modeled by the flow of a visual line—such as the mountain-range lines—across many unit cells.

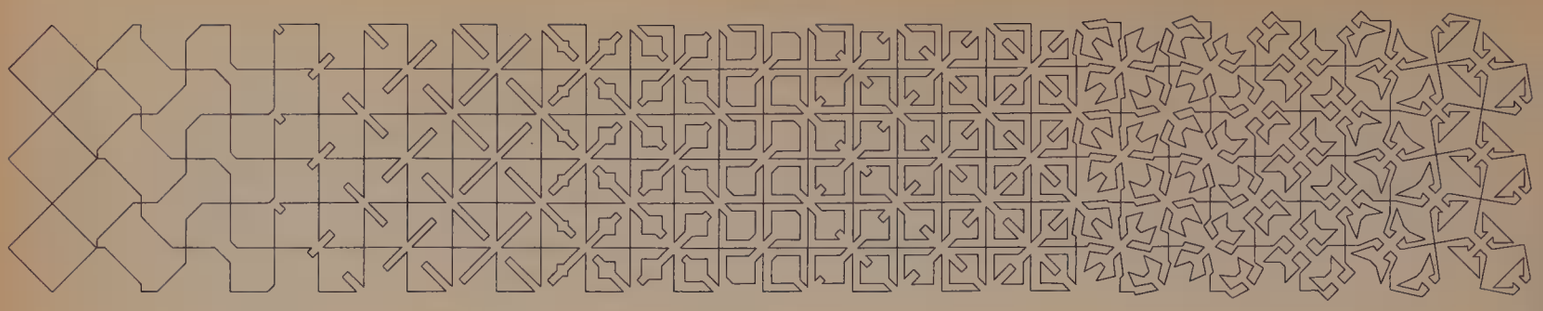
Bach's music is always called up in discussions of the relation between mathematical patterns and music, and this occasion is no exception. I am reminded particularly of some of Bach's texturally more uniform pieces, such as certain preludes from "The Well-tempered Clavier," where in each measure there is a certain pattern executed once or twice and possibly more times. From measure to measure this pattern undergoes a slow metamorphosis, meandering in the course of many measures from one region of harmonic space to far-distant regions and then slowly returning by some circuitous route. For specific examples you might listen to (or look at the scores of) Book I, No. 1 and No. 2, and Book II, No. 3 and No. 15. Many of the other preludes have this feature in places, although not for their entirety.

Bach seldom deliberately set out to play with the perceptual systems of his listeners. Artists of his century, although they occasionally played perceptual games, were considerably less sophisticated about, and less fascinated by, issues we now deem part of perceptual psychology. Such phenomena as regrouping would have intrigued Bach, and I sometimes wish he had known of certain effects and had been able to try them out, but then I remind myself that whatever time Bach might have spent playing with newfangled ideas would have had to be subtracted from his time for producing the masterpieces we know and love, and so why tamper with something that precious?

On the other hand, I do not find this argument 100 percent compelling. Who says that if you are going to imagine playing with the past, you have to hold the lifetimes of famous people constant in length? If we can imagine telling Bach about perceptual psychology, why can't we also imagine adding a few extra years to his lifetime to let him explore it? After all, the only *divinely* imposed (that is, absolutely unslippable) constraint on Bach's years is that they and Mozart's years add up to 100, no? Hence if we give Bach five extra years, then we merely take five away from Mozart. It is painful, to be sure, but not all *that* bad. We could even let Bach live to 100! (Mozart would never have existed.)

Although it is difficult to imagine and impossible to know what Bach's music would have been like if he had lived in the 20th century, it is certainly not impossible to know what Steve Reich's music would have been like if *he* had lived in this century. In fact, I am listening to a record of it right now. Now, Reich's music really *is* conscious of perceptual psychology. All the way through he plays with perceptual shifts and ambiguities, pivoting from one rhythm to another, from one harmonic origin to another, constantly keeping the listener on edge and tingling with nervous energy. Imagine a piece resembling Ravel's "Boléro," only with a much finer grain size, so that instead of its having roughly a one-minute unit cell it has a three-second unit cell. Its changes are so tiny that sometimes you can barely tell it is changing at all, whereas at other times the changes jump out at you. What Reich piece am I listening to? Well, it hardly matters, since most of his music satisfies this characterization, but for the sake of specificity you might try "Music for a Large Ensemble," "Octet" or "Violin Phase."

Let us now return to parquet deformations. "Dizzy Bee," executed by Richard Mesnik at Carnegie-Mellon in 1964 [*page 15*], involves perceptual tricks of another kind. The left side looks like a perfect honeycomb or, somewhat less poetically, a perfect bathroom floor. When we move to the right, its perfection seems in doubt as the rigidity of the lattice gives way to shapes that seem rounder. Then we notice that three of them have combined to form one larger shape: a superhexagon made up of three somewhat squashed pentagons. The curious thing is that if we now sweep our eyes from right to left back to the beginning, we can no longer see the left side in quite the way we



"Cucaracha"

saw it before. The small hexagons now are constantly grouping themselves into threes, although the grouping changes quickly. In our mind we experience "flickering clusters" where groups form for an instant and then disband, their components immediately regrouping in new combinations. The poetic term "flickering clusters" comes from a famous theory of how water molecules behave, in which the bonds are hydrogen bonds rather than mental ones.

Even more dizzying, perhaps, is "Consternation," executed by Scott Grady at SUNY at Buffalo in 1977 [*page 15*]. This is another parquet deformation in which hexagons and cubes vie for perceptual supremacy. It is so complex and agitated in appearance that I scarcely dare to attempt an analysis of it. In its intermediate regions I find the same exciting kind of visual pseudochaos as exists in Escher's best deformations.

Perhaps irrelevantly, but I suspect not, the names of many of these studies remind me of pieces by Zez Confrey, a composer best known in the 1920's for his novelty piano pieces such as "Dizzy Fingers" and "Kitten on the Keys" and—my favorite—"Flutter By, Butterfly." Confrey specialized in pushing rag music to its limits without losing musical charm, and some of the results seem to me to have a saucy, dazzling appeal not unlike the jazzy appearance of this parquet deformation.

The next parquet deformation, "Oddity out of Old Oriental Ornament," executed by Francis O'Donnell at Carnegie-Mellon in 1966 [*page 16*], is based on an extremely simple principle: the insertion of a "hinge" in one single line segment and the subsequent flexing of the segment at that hinge. The reason for the stunningly rich results is that the unit cell giving rise to the tessellation occurs both vertically and horizontally, so that flexing it one way induces a crosswise flexing as well, and the two flexings combine to yield this curious and unexpected pattern.

Another deformation that shows the amazing results of an extremely simple but carefully chosen transformation principle is "Y Knot," executed by Leland Chen at SUNY at Buffalo in 1977 [*page 16*]. If you look at it with full attention, you will see that its unit cell is in the shape of a three-bladed propeller and that the unit cell never changes in shape. All that does change is the *Y* lodged tightly inside the unit cell. And the only way the *Y* changes is by very slowly rotating clockwise. Admittedly in the final stages of rotation this forces some previously constant line segments to extend themselves a bit, but that does not change the outline of the unit cell in any way. It is remarkable what well-chosen simplicity can do.

Three of my favorites are "Crazy Cogs" (Arne Larson, Carnegie-Mellon, 1963), "Trifoliolate" (Glen Paris, Carnegie-Mellon, 1966) and "Arabesque" (Joel Napach, SUNY at Buffalo, 1979) [*pages 16 and 17*]. They all share the feature of getting increasingly intricate as you move to the right. Most of the preceding deformations do not have this extreme quality of irreversibility, that is, the ratcheted quality signaling that an evolutionary process is taking place. I can't help wondering if the designers did not think they had painted themselves into a corner, particularly in the case of "Arabesque." Is there any way you can back out of that supertangle except by retrograde motion, namely by retracing your steps? I suspect there is, but I wouldn't care to try to find it.

As a contrast consider "Razor Blades," an extended study in relative calmness [*preceding page*]. It is unsigned but was executed at Carnegie-Mellon in 1966. Like "Fylfot Flipflop," the first piece I described, this one can be broken up into long, wavy horizontal lines and vertical structures crossing them. It is a little easier to see them if you start at the right side. For instance, you can see that just below the top there is a long snaky line with numerous little nicks in it, undulating its way to the left and in so doing shedding some of those nicks, so that at the very edge it has degenerated into a perfect square wave, as such a periodic waveform is called in Fourier analysis. Complementing this horizontal structure is a similar vertical structure that is harder to describe. The thought that comes to my mind is that of two ornate, rather rectangular hourglasses with ringed necks, one on top of the other. You can see for yourself.

As with "Fylfot Flipflop," each of these patterns by itself is intriguing, but of course the real excitement comes from the daring act of superposing them. Incidentally, I know of no piece of visual art that better captures the feeling of beauty and intricacy in a Steve Reich piece, created by slow "adiabatic" changes floating on top of the chaos and dynamism of the lower-level frenzy. Looking back, I see I began by describing this parquet deformation as "calm." Well, what do you know? Perhaps I would be a good candidate for one of *The New Yorker*'s occasional notes titled "Our Forgetful Authors."

More seriously, there is a reason for this inconsistency. One's emotional response to a given work of art, whether the work is visual or musical, is not static and unchanging. There is no way of knowing how you will respond the next time you hear or see one of your favorite pieces. It may leave you unmoved or it may thrill you to the bone. It depends on your mood, on what has recently happened, on what happens to strike you and on many other subtle intangibles. One's reaction can even change in the course of a few minutes. And so I won't apologize for this seeming lapse.

Let us now look at "Cucaracha," executed by Jorge Gutierrez at SUNY at Buffalo in 1977 [*above*]. It moves from the utmost geometricity—a lattice of perfect diamonds—through a sequence of gradually more arbitrary modifications until it reaches some kind of near-freedom, a dance of strange, angular, quasi-organic forms. This fascinates me. Is entropy increasing or decreasing in this rightward flow toward freedom?

A gracefully spiky deformation is the one wittily titled "Beecombing Blossoms," executed by Laird Pylkas at SUNY at Buffalo this year [*opposite page*]. Huff told me Pylkas struggled for weeks with it and at the end, when she had resolved her difficulties, she mused: "Why is it that the obvious ideas always take so long to discover?"

As our last study let us take "Clearing the Thicket," executed by Vincent Marlowe at SUNY at Buffalo in 1979, which mixes straight lines and curves, right angles and cusps, explicit squarish swastikoids and implicit circular holes [*page 20*]. Rather than demonstrating

my inability to analyze the ferocious complexity of the design, I should like to use it as the jumping-off point for a brief discussion of computers and creativity.

Some totally new things are going on in this particular parquet deformation, things that have not appeared in any previous one. Notice the hollow circles at the left that shrink as you move to the right; notice also that at the right there are hollow "anticircles" (concave shapes made from four circular arcs turned inside out) that shrink as you move to the left. Now, according to Huff, such an idea had not appeared in any previously created deformations. This means that something unusual happened here: something genuinely creative, something unexpected, unpredictable, surprising, intriguing and, not least, inspiring to future creators.

Thus the question naturally arises: Would a computer have been able to invent this parquet deformation? Well, put this way it is a naive and ill-posed question, but we can try to make some sense of it. The first thing to remind ourselves of is that the term "computer" refers to nothing more than an inert hunk of metal and semiconductors. To go along with this bare computer, this hardware, we need software and energy. The former is a specific pattern inserted into the hardware, binding it with constraints and yet imbuing it with goals; the latter is what breathes "life" into it, making it act according to those constraints and goals.

The next point is that the software is what really controls what the machine does; the hardware simply obeys the software's dictates, step by step. And yet the software could exist in a number of different "instantiations": realizations in different computer languages. What really counts about the software is not its literal aspect but a more abstract, general, overall "architecture," which is best described in a nonformal language such as English. We might say that the plan, the sketch, the central idea of a program is what we are talking about here, not its final realization in some specific formal language or dialect. That is something we can leave to apprentices to carry out after we have presented them with our informal sketch.

Hence the question actually becomes less mundane and more theoretical and philosophical: Is there an architecture to creativity? Is there a plan, a scheme, a set of principles that, if it were elucidated clearly, could account for all the creativity embodied in the collection of all parquet deformations, past, present and future?

Note that we are asking about the *collection* of parquet deformations, not about some *specific* work. It is a truism that any specific work of art can be re-created, even re-created in various slightly novel ways, by a programmed computer.

For example, Piet Mondrian evolved a highly idiosyncratic, somewhat cryptic style of painting over a period of many years. You can see, if you trace his development over time, exactly where he came from and where he was headed. If you focus on a single Mondrian work, however, you cannot sense this stylistic momentum, this quality of dynamic, evolving style that any great artist has. Looking at one work in isolation is like taking a snapshot of something in motion: you capture its instantaneous position but not its momentum. Of course, the snapshot might be blurred somehow, in which case you would get a sense of the momentum but would lose information about the position. When you are looking at just a single work of art, however, there is no mental blurring of its style with that of recent works or soon-to-come works; you have exact position information ("What is the style *now*?") but no momentum information ("Where was the style and where is it going?").

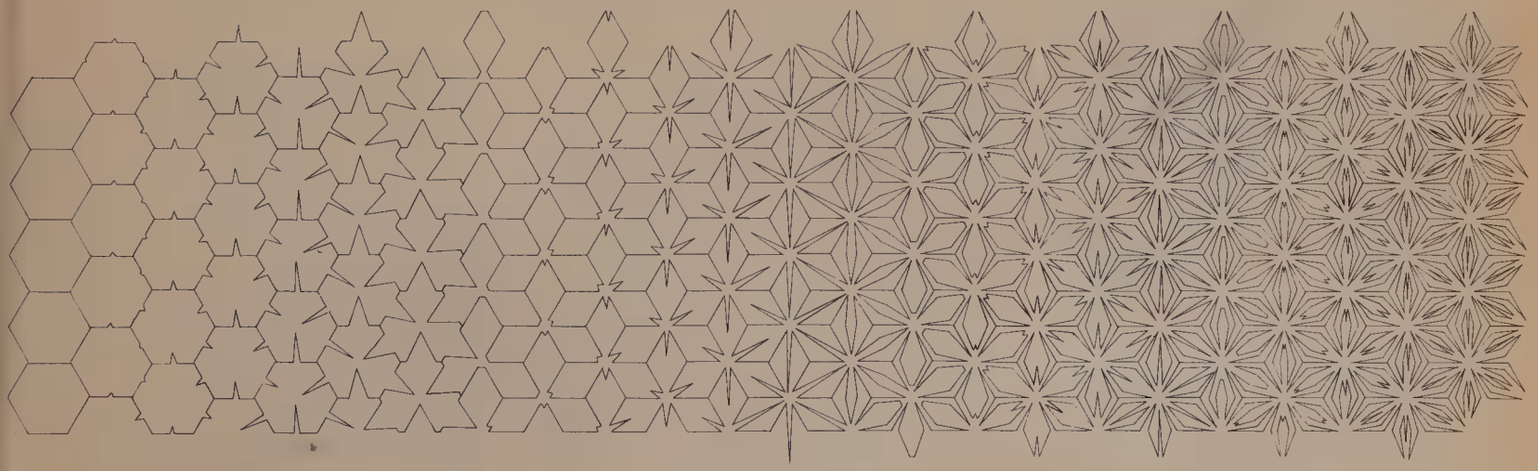
Some years ago A. Michael Noll, a mathematician and computer artist, took a single Mondrian painting—an abstract, geometric study with seemingly random elements—and from it extracted some statistics on the patterns. Given these statistics, he programmed a computer to generate numerous pseudo-Mondrian paintings having the same or different values of these randomness-governing parameters. Then he showed the results to viewers who had no foreknowledge of what he was up to. The reactions were interesting: more people preferred one of the pseudo-Mondrians than preferred the genuine Mondrian!

This is quite amusing, even provocative, but it also is a warning. It proves that a computer can certainly be programmed, after the fact, to imitate—and imitate well—mathematically capturable stylistic aspects of a given work. It also warns us, however: Beware of cheap imitations!

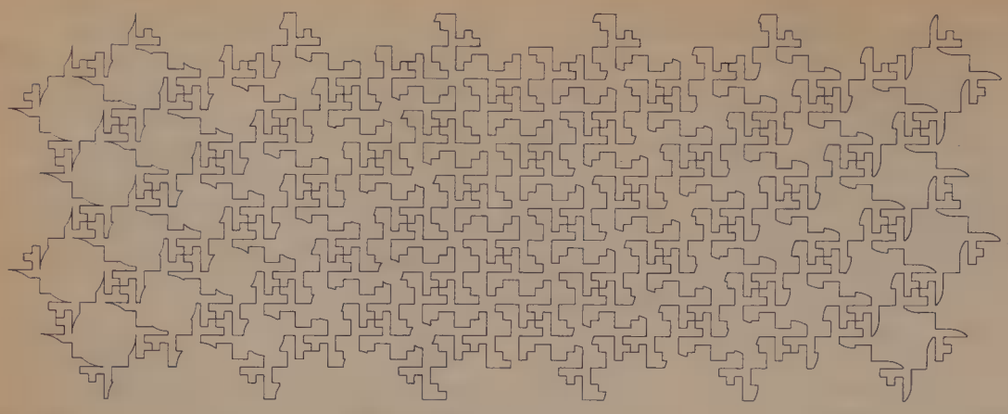
Consider parquet deformations. Undoubtedly a computer could be programmed to do any *specific* parquet deformation or minor variations on it without too much trouble. There simply are not that many parameters to any given deformation. The essence of any artistic act lies not, however, in selecting particular values for certain parameters but far deeper: in the balancing of myriad intangible and mostly unconscious mental forces, a judgmental act that results in many conceptual choices eventually adding up to a tangible, perceptible, measurable work of art.

Once the finished work exists scholars looking at it may seize on certain qualities of it that lend themselves to being easily parametrized. *Anyone* can do statistics on a work of art once the work is there for the scrutiny, but the ease of doing so can obscure the fact that no one could have said a priori what kinds of mathematical observables would turn out to be relevant to the capturing of stylistic aspects of the as-yet-unseen work of art.

Huff's own view on this question of mechanizing the art of parquet de-



"Beecombing Blossoms"



"Clearing the Thicket"

formations closely parallels mine. He believes some basic principles could be formulated at the present time that would enable a computer to generate relatively stereotyped yet novel creations of its own. He stresses, however, that his students occasionally come up with rule-breaking ideas that enchant the eye for reasons deeper than any he has been able to verbalize. In this way the set of explicit rules gets gradually enlarged.

Comparing the creativity that goes into parquet deformations with the creativity of a great musician, Huff writes: "I don't know about the consistency of the genius of Bach, but I did work with the great American architect Louis Kahn (1901–1974) and suppose it must have been somewhat the same with Bach. That is, Kahn, out of moral, spiritual and philosophical considerations, formulated ways he would and ways he would not do a thing in architecture. Students came to know many of his ways, and some of the best could imitate him rather well (although not perfectly). But as Kahn himself developed he constantly brought in new principles that brought new transformations to his work, and he even occasionally discarded an old rule. Consequently he was always several steps ahead of his imitators who knew what *was* but couldn't imagine what *will be*. And so it is that computer-generated 'original' Bach is an interesting exercise. But it isn't Bach—that unwritten work that Bach never got to, the day after he died."

The real question is: What kind of architecture is responsible for *all* these ideas? Or is there any *one* architecture that could come up with them all? I would say that the ability to design good parquet deformations is probably deceptive in the same way as the ability to play good chess is: it seems more mathematical than it actually is.

A brilliant chess move, once the game is over and can be viewed in retrospect, can be seen as logical, as "the correct thing to do in that situation." But brilliant moves do not originate from the kind of logical analysis that occurs after the game; there is no time during the game to check out all the logical consequences of a move. Good chess moves spring from the organization of a good chess mind: a set of perceptions arranged in such a way that certain kinds of ideas leap to mind when subtle patterns or cues are present. The way perceptions have of triggering old and buried memories underlies skill in any type of human activity, not only chess. It is just that in chess the skill is particularly deceptive, because after the fact it can all be justified by a logical analysis, a fact that seems to hint the original idea *came from* logic.

Writing lovely melodies is another one of those deceptive arts. To the mathematically inclined, notes seem like numbers and melodies like number patterns. Therefore all the beauty of a melody *seems* as if it ought to be describable in some simple mathematical way. So far, however, no formula has produced even a single good melody. Of course, you can look back at any melody and write a formula that will produce it and variations on it. But this is retrospective, not prospective. Lovely chess moves and lovely melodies (and lovely theorems in mathematics) have this in common: every one has idiosyncratic nuances that seem logical a posteriori but are not easy to anticipate. To the mathematical mind chess-playing skill and melody-writing skill and theorem-writing skill seem obviously formalizable, but the truth turns out to be more tantalizingly complex than that. Too many subtle balances are involved.

So it is with parquet deformation, I reckon. Each one taken alone is in some sense mathematical. Taken as a class, however, they are not mathematical. This is what is tricky about them. Don't let the apparently mathematical nature of an *individual* deformation fool you; the architecture of a program that could create all these parquet deformations and more good ones would have to incorporate computerized versions of *concepts* and *judgments*, and those are much more elusive and complex than mere numbers.

At this point many critics of computers and artificial intelligence, eager to find something "computers can't do" (and never will be able to do), often go too far: they jump to the conclusion that art and, more generally, creativity are fundamentally uncomputerizable. This is hardly the implied conclusion! The implied conclusion is that if computers are to be enabled to act human, we shall have to wait until we have good computer models of such human properties as perception, memory, mental categories, learning and so on. We are a long way from that. There is no reason to assume, however, that those goals are in principle unattainable, even if they remain far off for a long time.

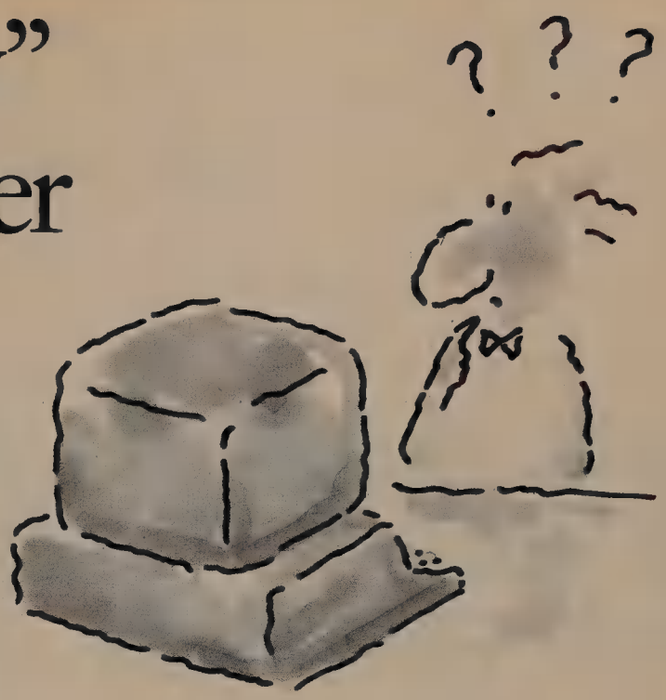
In this column I have been playing with the double meaning of the term "architecture": it means both the design of a habitat and the abstract essence of a grand structure of any kind. The former has to do with hardware and the latter with software. In a certain sense, William Huff is a professor of both brands of architecture. Obviously his professional training is in the design of "hardware," namely genuine habitats for human beings, and he works in a school where that is what they do. He is also, however, in the business of forming in the minds of his students a softer kind of architecture: the mental architecture that underlies the skill to create beauty. Fortunately for him he can take for granted the complexity of a human brain as the starting point on which to build this architecture.

When I first met Huff and saw how abstract and seemingly impractical were the marvelous works produced in his design studio—ranging from parquet deformations to strange ways of slicing a cube to gestalt studies using thousands of dots to eye-boggling color patterns—I at first wondered why this man was a professor of *architecture*. But after talking with him and his colleagues my horizons were extended about the nature of their discipline.

The architect Louis Kahn had great respect for the work of William Huff, and it is with his words that I should like to conclude:

"What Huff teaches is not merely what he has learned from someone else but what is drawn from his natural gifts and belief in their truth and value. In my belief what he teaches is the introduction to discipline underlying shapes and rhythms, which touches the arts of sight, the arts of sound and the arts of structure. It teaches students of drawing to search for the abstract and not the representational. This is good as a reminder of order for the instructors/architectural sketchers (like me), and good especially for the student sketchers without background. It is the introduction to exactitudes of the kind that instill the religion of the ordered path."

Take "Know" for an Answer



Read *PC World*, The Most Indispensable New Magazine for Those Who Want to Know the Bottom Line About the IBM Personal Computer and IBM PC Compatibles

Are you now using or do you plan to use an IBM Personal Computer or an IBM-compatible personal computer?

Does it all seem very confusing to you?

If you don't know who to turn to for the most objective and impartial advice on the subject, you can easily get turned inside out. Or worse still, you can get turned off.

Now you can refer to *PC World*, the newest, liveliest, and most authoritative magazine, devoted exclusively to IBM and IBM-compatible personal computer users.

With its visually stunning format, colorful illustrations, and easy to read articles, *PC World* is designed to appeal to both technical and nontechnical readers.

Every month *PC World* will keep you posted on the most interesting, intriguing, and potentially useful applications for this remarkable machine.

It not only tells you about the best products on the market, but also how to use them.

In fact, *PC World* is one of the most inexpensive peripherals you can buy for your personal computer; it allows you to get the maximum use of your PC for a minimum investment.

Now for a limited time only you can become a Charter Subscriber to a magazine that's destined to reach the top of the personal computer charts. Enter your subscription today using the coupon and save up to 56% off the newsstand price.

Now *that's* the bottom line.

Yes, I'll take "know" for an answer.
Please enter my subscription to
PC World at a special introductory rate!

Name_____

Company_____

Address_____

City/State/Zip_____

U.S. & Canada only
☐ 12 issues $18.95 (47% savings)
☐ 24 issues $33.50 (53% savings)
☐ 36 issues $47.00 (56% savings)
☐ payment enclosed ☐ bill me later
☐ Visa ☐ MasterCard
expiration date_____
card #_____
Interbank # (MC only)_____

Toll-Free: **800/247-5470**
from Iowa 800/532-1272

Please allow 6–8 weeks for delivery.

PC World
Subscription Dept.
P.O. Box 6700
Bergenfield, NJ 07621



M0147

BOOKS

Cholera, whales, statistical snapshots of the U.S., rare wonders of the Incas

by Philip Morrison

CHOLERA: THE AMERICAN SCIENTIFIC EXPERIENCE, 1947–1980, by W. E. van Heyningen and John R. Seal. Westview Press, Boulder, Colo. ($27.50). A thousand miles of the leeward southern slopes of the Himalayas receive the fierce monsoon rains, to drain into the sea through the many mouths of the broad Ganges. That intricate delta is the heart of Bengal, populous and fertile, 200 miles across between its two capitals east and west, Dacca and Calcutta. For a long time the dwellers in those wetlands have endured an endemic bacterial disease, cholera, that nearly every year takes a high toll in human life. It is an abrupt and terrifying affliction, "the bowels turning to water"; the patient—chilled, thirsty, vomiting as he passes a clear, watery stool by the pint every hour or two, yet alert—soon develops an "intense fear of impending death." That fear is justified within a few days, for as many as two cases out of three.

Two centuries ago, perhaps following some genetic change in the cell strain, cholera learned to travel. Six pandemics followed; the calamity became a catchword and a curse throughout the West. Americans counted their dead from cholera by the hundred thousand as recently as the years after the Civil War. During the worst visitation of Britain, in the 1850's, the queen's "incomparable" physician, John Snow, demonstrated beyond doubt that the infection was carried in water contaminated by sewage. He did not even need to know that its cause was a comma-shaped bacillus, found swarming in the intestinal contents of victims in the same years by Filippo Pacini. In capital-rich countries the spread of *Vibrio cholerae* has for most of a century been avoided by the scrupulous separation of piped drinking water from sewage, although it took a generation before Snow's proof led to public action. But the world's resources are so ill-shared that most tropical countries cannot yet afford such prevention for the million villages at risk.

This engrossing and varied volume, as much at home in the laboratory politics of medical expatriates and their capricious institutional patrons as it is in cyclase biochemistry, brings the story up to date. Cholera is expensive to prevent, although the developed world has managed it. Over the past decade, however, it has become cheaply curable, through insight. Two authors, the first an Oxford biochemist, the second an American medical-research administrator, are veterans of the campaign. A 30-year war against the vibrio has been waged across the world by a shifting international team of men and women, largely Americans who knew little of cholera but much of physiology and biochemistry, allied with the experienced doctors of the cholera countries.

Deadly cholera is a superficial disease, only skin-deep. The thin layer of cells affected is the epithelial lining of the small intestine, an area measuring 200 square meters, a broad living carpet "fit for the Palace of Versailles," all cunningly folded into the 10-foot tube of the small intestine packed within each human abdomen. Only a single layer of cells is affected by the microbes, which secrete a specific protein toxin. The toxin quickly enters and raises the level of the messenger substance (cyclic adenosine monophosphate) in the cells, in analogy to cell response to hormones. This intricate piece of biochemistry is quite well understood, although just how the message changes the cell is not yet known.

The result is a shift in the balance between the absorption and the secretion of water and its dissolved ions by that cell layer. Each hour the normal intestinal wall passes four or five liters of water; ordinarily the balance inward and outward is elegantly struck. Only a few minutes' flow is needed to replace the normal net loss over a day. A photograph shows a slight, cheerful man, a newly recovered cholera patient, "surrounded by 96 litre bottles" that had held the intravenous fluid that pulled him through a few days' illness. A loss of 30 quarts of water a day is not untypical. The disease affects only that thin cell layer; the toxin never enters the bloodstream.

Those cells are in any case sloughed off steadily from the gut; in normal circumstances they live and die in less than a week. The toxin, however, poisons them at once; they simply remain without adequate function for a few days, until they are duly replaced by unaffected new cells in the usual sequence. The disease is therefore quickly self-limiting, less systemic than a common cold. The patient dies, but not from the toxin, which has only induced an invisible biochemical lesion in that single layer of short-lived cells. Death is from intense dehydration and demineralization, the lifeblood turned dark and thick. Early workers tried to relieve the symptoms by giving the patient a lot of water to drink, but their efforts worked only transiently; an hour or two later the shadow of death was back. Enough water just would not go in.

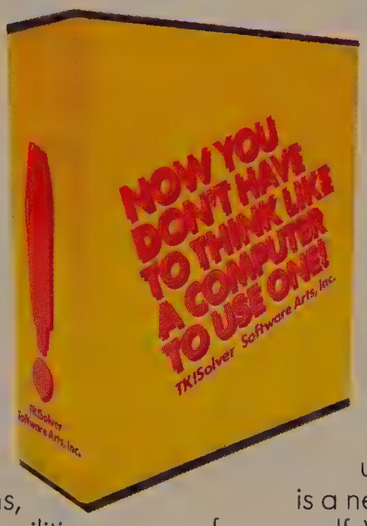
It was in the fall of 1947, during a rogue outbreak of cholera along the Nile, that the nature of death from cholera was grasped. An epidemiological team concluded from its surprise encounter with the acute disease that "the diarrhoea . . . constitutes the entire pathologic aspect of the disease." That team was a new U.S. Navy group, its head a Rockefeller Institute physiologist, Comdr. Robert Allan Phillips, M.C., U.S.N. Phillips had come to wartime Cairo as a naval medical officer to join the U.S. Typhus Commission, set up there in 1942 to unite American, British and Egyptian efforts to deal with typhus, feared in North Africa both during and after the war.

Phillips had a taste for the role, and so once the war ended and his Cairo laboratory was about to be disbanded he "rushed back to Washington and made his pitch for support to every likely agency." Only the Navy, with its long tradition of concern for tropical disease and a flexible and enthusiastic postwar view of research, acted. Cholera attacked almost at once; the unit's improvised but brilliant defense marked the beginning of the modern American campaign against cholera, led over the years by a shifting set of delicate alliances among "the Navy, the Army, the Department of State, the Department of Health, Education, and Welfare, and the academic world."

This chronicle is as diverse as the story of any modern war. The politics of the committee room late at night in some Atlantic City hotel can fix or move forces halfway across the world; foreign policy opens and closes research. Field leaders are removed or promoted as at home their friends and critics clash or concur. In the field changing ideas and rapid actions count, yet chance shapes results. Directors matter, but so does the talent of Mr. Razzack, the mechanic. He kept at work the 37 motor launches whose river ambulance patrols in the

CRUNCH!

The inventor of VisiCalc,® the world's leading software product, introduces TK!Solver™—the revolutionary way to crunch problems on a personal computer.



With VisiCalc, Software Arts™ created an entirely new application for personal computers—the electronic spreadsheet.

Now Software Arts has developed TK!Solver, another entirely new and different application for personal computers—the problem solver. Business and financial problems. Scientific and engineering problems. Architectural and building design problems. Any kind of problem you can dream up, in any area.

What's so revolutionary about that? Simply this: the power of the TK!Solver program comes from the ease with which you can set up problems, vary assumptions, find solutions, and display results. All the facilities needed to solve the problems—simple or complex—are built-in, and need not be developed. Once you have stated your problem, simply enter the known values, then solve the problem with a single keystroke.

For even easier problem solving use a TK!SolverPack™ in which models for solving common problems in such areas as finance, engineering and education have already been formulated by professionals in these fields.

There is no software product in existence that solves problems as quickly, flexibly, and easily as TK!Solver—The Problem Cruncher™. In fact, no other software product even works the way TK!Solver does. It is a necessity for personal computer owners. See for yourself. Visit your nearest computer store for a demonstration and find out how much you can do with a personal computer when it thinks as flexibly as you do.

Software Arts™

27 Mica Lane, Wellesley, MA 02181

TK!Solver, TK!SolverPack and TK!Solver—The Problem Cruncher are trademarks of Software Arts, Inc. Software Arts is a trademark of Software Arts, Inc. and Software Arts Products Corp. VisiCalc is a trademark of VisiCorp. Copyright © 1983. Software Arts, Inc. All rights reserved.

cholera season along the waterways a couple of hours out of Dacca by boat were the chief means of the field trial of cholera vaccine there in 1964. So does the courage of the Baltimore volunteers who earlier tested the vaccine.

The battles and the medical units are too many to list here. The first decade and more were spent in learning how to save cholera patients by the prodigal intravenous injection of sterile, isotonic, pyrogen-free saline water under knowing control. In Thailand, in the Philippines, in South Vietnam, wherever the teams were at work, cholera took less than 1 percent of those who came under the painstaking care required for injecting floods of intravenous saline. But few cholera patients could come under such care, which demanded at least a large corps of trained medical aides, if not doctors and hospitals. The defeat of vaccines—both those using cells and those using the newly crystallized toxin—has been disappointing, although not unexpected in a disease that injures so quickly and so superficially. The cheap antibiotic tetracycline kills the vibrios, but only after their thin-layer damage is already done.

The simple cure, almost if not quite a final victory, came in about 1968, presaged by an unhappy but revealing trial under Phillips in Manila in 1962 and by various laboratory results even earlier. The Dacca and Calcutta centers both confirmed about then that adding glucose to drinking water enabled the patient to absorb enough water; the rationale appears to be that the passage of a triple complex of sodium ion, glucose molecule and water molecule through the decisive membrane is not at all inhibited by the cholera toxin. Field trials were forced by history on the Calcutta laboratory run by Johns Hopkins in the monsoon months of 1971, when six million refugees fled westward from the civil war in East Pakistan. Triumph! Nearly 4,000 patients were treated, with an overall fatality rate under 4 percent. The treatment was oral fluids made up of local materials packaged on the spot in the laboratory's library, enough for all the patients at a total cost of $750.

The victory is not quite total; it may be that some people will always need intravenous rehydration and not just those cheap cupfuls of sugared and salted drinking water. But it looks as though the toxin of cholera, as deadly as that of tetanus or botulinus but specific to epithelial cells instead of to the deep-lying



GIVE YOURSELF A WHOLE NEW SLANT ON PERFORMANCE—A TURBO SPORTS CO

nerve endings of the central or motor system, has been routed. The defeat is not by medical prevention—unless the piped water and adequate nutrition of the well-to-do can be called medical—but by cure, a simple cure by drinking water in the right way.

There is a new center now in Dacca, internationally supported and administered but at home there, dedicated by the president of Bangladesh in 1979. It has high hopes and a wide brief: no less than global attack on all the diarrheal diseases, together accounting for 500 million cases each year, a chief cause of death in children particularly. In 1978 *The Lancet* wrote: "The discovery that sodium transport and glucose transport are coupled in the small intestine, so that glucose accelerates absorption of solute and water, was potentially the most important medical advance this century." The total dollar cost to U.S. taxpayers of their generation-long contribution to the war against cholera was about the cost of one of today's air-superiority fighter planes.

MOBY-DICK; OR, THE WHALE, by Herman Melville. Illustrated with 100 woodcuts by Barry Moser. University of California Press ($25). THE ECOLOGY OF WHALES AND DOLPHINS, by D. E. Gaskin. Heinemann Educational Books, Inc. Exeter, N.H. ($45). Does the whale's magnitude diminish? Will he perish? In Chapter 105 of his ocean of a book Melville asks those questions and then answers them: "And as upon the invasion of their valleys, the frosty Swiss have retreated to their mountains; so, hunted from the savannahs and glades of the middle seas, the whalebone whales can at last resort to their polar citadels... and bid defiance to all pursuit from man.... Wherefore we account the whale immortal in his species." Melville wrote of three-masted ships, before Svend Foyn and his cannon with its grenade-tipped harpoon, before the powered catcher boat and the big floating factories. This is no place to appraise Melville's own prose leviathan; we only note here a typographically beautiful edition of the work, decorated with striking woodcuts that avoid visual interpretation of character or event, to present only the tools, the prey, the ships and the ports. The handsome folio was hand-set in Goudy Modern and hand-printed by Andrew Hoyem in 1979 in a few hundred copies; this edition is a reduced facsimile of that collector's plea-



SAFETY IS A SNAP
American Motors
Renault Fuego Turbo. Fuel-injected intercooled turbo with 11.6 psi maximum boost. 5-speed overdrive transmission. Front wheel drive. Power-assisted front discs. Michelin TRX radials on cast aluminum wheels. Air conditioning. AM/FM stereo radio with quad speakers. Tach-to-turbo gauge telemetry. Leather-wrapped steering wheel, and more. $11,095* *List price. Tax, license, destination charges, optional or regionally required equipment extra
OM THE PEOPLE WHO TURBOCHARGED FORMULA ONE RACING.
RENAULT
THE ONE TO WATCH



It's now possible for people who've only admired great pictures to take them.



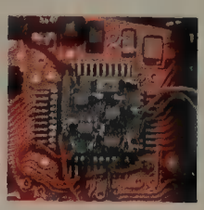
The Programmed FG.

Inside everyone who's ever gazed appreciatively at a photograph is the urge to take great pictures themselves.

And now, no matter how much or how little picture taking experience you've had, there's a camera worthy of an admiring look.

The remarkable programmed FG.

The FG is a camera that ambitious beginners can still use after they've become accomplished veterans.



In the programmed mode, just focus and shoot. A microchip computer automatically selects the proper lens opening and shutter speed.

And through-the-lens flash metering with the optional SB-15 speedlight makes flash pictures virtually fool-proof.

Automatic and manual, too.

But what happens when you want a camera that allows you more flexibility?

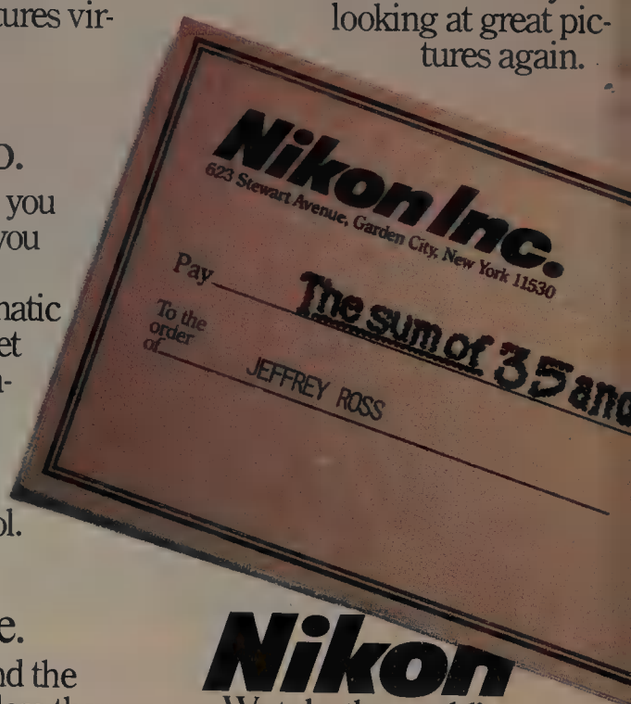
Then the FG is an automatic camera that allows you to set the aperture yourself to control depth-of-field. And, in the manual mode, you can make all settings yourself for complete creative control.

$35 rebate for a limited time.

But the real genius behind the FG is that it costs so much less than you'd think. And until August 15, 1983, there is a $35 cash rebate.*

So visit your Nikon camera dealer soon and check out the ingenious programmed FG.

You'll never be content just looking at great pictures again.



Nikon
We take the world's greatest pictures.™

*Offer applies only to products which include the Nikon, Inc. limited U.S.A. warranty form. Photographs courtesy of the Witkin Gallery, Inc., N.Y.C. © Nikon Inc. 1983. For further information write Dept. 47, Garden City, N.Y. 11530.

sure, down in page area by a factor of about two and in price by one of 40.

D. E. Gaskin addresses the very question Melville set. He is a marine biologist at the University of Guelph in Ontario. He counts 20 years in the study of whales and their kin, with field experience from the circumpolar Southern Ocean to the Newfoundland banks. His volume is a critical technical review of what is known—and not known—about the main species of cetaceans, looked at from the viewpoint of populations over time and space, with only the indispensable minimum of material on the individual animals and their adaptations. As Melville's whale was metaphysical, so is our whale today political, "a semi-mythological beast."

A summary of such a thorough review of an exponentially growing scholarly literature is perforce schematic. In the first third of the book whales are presented as seasonably migratory animals (several species are discussed in detail) that find distinct requirements for breeding and for the taking of food; their distribution is a long-evolved compromise between these two needs. Their travels are empathetic: winter in the sun, summer in the cool. It is the prey that leads the whale; sperm whales take mostly squid in the open sea, mostly fish closer to shore. Sperm migrations—recall the white whale—are diffuse, ill-defined. (Although the big males are locally overexploited, this wide-ranging species numbers some 600,000 worldwide: "a satisfactory condition.") But the decimated plankton-feeding giants, the blues, move and group regularly from Tropic to pole with the sun because their polar food, the krill, are particularly seasonal, patchy and discontinuous.

How whales navigate and even how they hunt we do not really know; echolocation, star sights, even shore landmarks may play a part for one species or another. A recent experiment with a single humpback, temporarily captive, found that the blindfolded animal could not run a simple maze of obstacles. It is such doubt, raised by experience, of the easy and attractive assumptions that follow from our modest insight into the animals that sets the tone of Professor Gaskin's cool book.

The metabolism of whales is of course a key for any fundamental ecological study. In one very interesting chapter it is made clear that whales and porpoises probably do not much differ from other mammals in their energy needs, the weight taken into account. An estimate for a 50-ton male cachalot, such as Moby-Dick, suggests an intake of squid at 3 percent of the whale's body weight per day; the female sperm whale, smaller by a factor of four or so, has a narrow margin of nutrition, particularly in view of the demands of lactation. This species lives usually segregated by sex; perhaps—it is a conjecture—that is because the energy demands of the two sexes are so different.

The big blues are biggest in the Antarctic; there the single-species patches of krill abound for only four months. In the Northern Hemisphere the fewer, smaller blues take more species of the zooplankton over a six-month season. The large size of the blue may be an adaptation to the short feeding season; the energetic issue is between staying on in a marginal feeding ground or cruising off to warmer waters. The matter is more complex still, because it appears (there is some doubt) that the big rorquals are too dense to float and must spend power just to stay near the surface to breathe. The management question hidden in all the careful budgets for propulsion, heat loss by conduction and radiation, and the like, is the retrospective calculation of how much krill the blues ate when they were numerous. The guesses suggest that the standing crop of krill was 10 or 20 times the total annual catch of all the world's fisheries. If the little krill live for a few years, as studies show, the future commercial krill take might be indefinitely sustainable at a level comparable to all the fishing the world now does for all species. Maybe.

Whales are gregarious to a degree. This text is rather skeptical about social implications, finding that simple groups and mutual finds of food are as likely as more integrated and complex behavior. It seems clear that groups of sperm whales keep in close acoustic-ranging contact as they forage in the depths. The animals disperse on diving but signal their identity frequently; each knows where the others are. They can return to the surface in a tight school; the calves, unable to dive as deep, may be protected by this stratagem from sharks.

A quite technical and uncertain third of the book seeks to understand the evolution and speciation of the order of whales and porpoises. Recent results on chromosome structure support the idea that all the cetaceans, small and large alike, came from a single line in the Paleocene, a view not popular until now.

Finally the author takes up the analysis and management of whale populations, the core of Melville's concern and our own. It is not easy to count whales: they are not randomly scattered, markers are hard to plant and to recover, age determination and recognition of individuals by their natural patterns are less than reliable. In 1980 a finback whale was for the first time fitted with a radio transmitter that survived; the animal was followed for 1,000 miles. These years we are passing a phase in animal ecology "largely as a result of intense love affairs with high-speed computing" that has plenty of models and rather less input data. Paper whales are simulated well with silicon chips, real ones and their hunters much less so.

Models of single species, like the lumped quotas and regulations dear to international agreements, are abstractions that may prove dangerously misleading. The chapter on the sorry history of postwar pelagic whaling is a gloomy one; Gaskin inclines to the view that mere short-run economics has been used to justify the near-extinction of species. The profitable idea is to match the lifetime of the gear to that of the fishery! The fishery declines sadly; years later there is a new big crop and new fishing investment: call it pulse fishing. Twenty years of pure research would not be a bad international plan for the present. Even though our analysis is poor, it remains true that the major failures have been ethical ones. The regime of local 200-mile limits, the lure of the krill, climatic effects on grain crops—all suggest that the diminished whales may not rest safe even in their lonely citadel.

U/S: A Statistical Portrait of the American People, edited by Andrew Hacker, with the assistance of Lorrie Millman. The Viking Press and Penguin Books ($25). Statistical Abstract of the United States: National Data Book and Guide to Sources, 103rd Edition, 1982–1983, U.S. Department of Commerce, U.S. Bureau of the Census. Superintendent of Documents, U.S. Government Printing Office ($15). The map is clear: since 1790 the point that marks the decennially reckoned centroid of the population of the U.S. has moved westward from the shores of Chesapeake Bay to vault the broad Mississippi for the first time in 1980. Ever since 1910 it has drifted in a straight line with a southerly trend, pulled by that distant southwestern desert sun.

The constitutional head count fixed the number of men, women and children resident in these United States on April 1, 1980, at just 226,545,805. This angelic accuracy is a fiction; the Bureau of the Census reckoned before the fact that it would be about 222 million people. The bureau received in the end about 2.2 percent more official name forms than that, a result seeming to mean that the undercounting by 2.5 percent the bureau itself had corrected for in 1970 had ceased. By 1982, however, the bureau, under suit in the courts, reversed its views. Black Americans were still undercounted, the bureau conceded, by nearly 5 percent. The report remains officially unchanged, although a figure near 228 million might be held a better approximation.

The decennial count is not the only source of these data today. Each month, mainly seeking the unemployment rate, the bureau questions a careful sample of



Another QUESTAR® . . .totally new. . . the long-range microscope

This is the QM1, the new instrument by Questar Corporation. It focuses at distances from 22 to 77 inches from the target. With it you will inspect hot materials at safe distances, observe live specimens in natural habitat, inspect micro-circuits, use it for crystal growth studies and a thousand other applications. It will let you scrutinize and photograph living organisms or radio-active samples with no damage to either one or yourself. And the magnificent design of these optics permits resolution below 2.5 microns. With a selection of eyepieces and extensions, magnifications can range as high as 150X.

Shown below on the right is a chip from a micro-circuit photographed with a normal 50mm lens to show its tiny size. On the left is a QM1 photograph of the lower right corner, reproduced from an 8 × 10 enlargement of 35mm Kodak Technical Pan. Distance 22 inches.

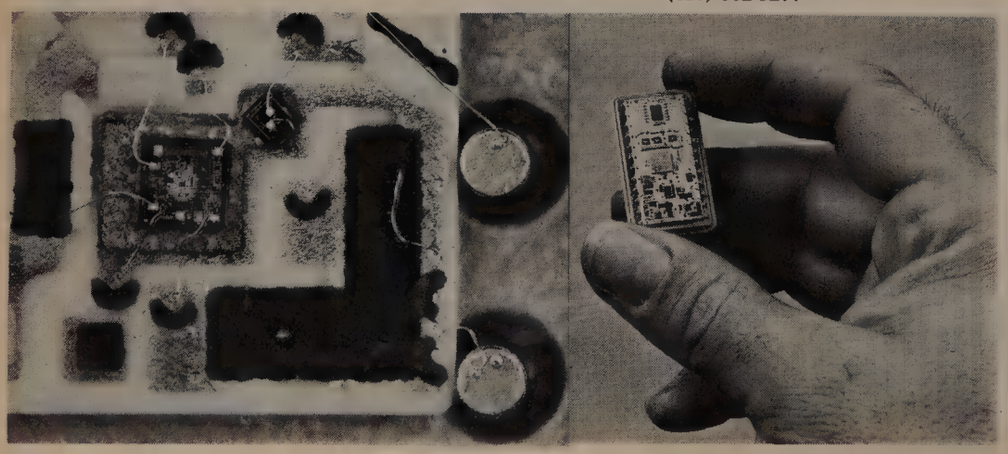
With the Questar QM1 we can provide full support systems of equipment and accessories, including video and high-speed motion cameras, lighting equipment, micro measurement and image analysis, and computer storage and display units. These systems are individually assembled to meet the customer's requirements.

NEW LITERATURE WITH SPECIFICATIONS AND PHOTOGRAPHS OF QM1 PERFORMANCE ON REQUEST.

The QM1: ***non-destructive testing, non-disruptive observation, non-contact gauging.***

QUESTAR

P.O. Box C, Dept. 205, New Hope, PA 18938
(215) 862-5277

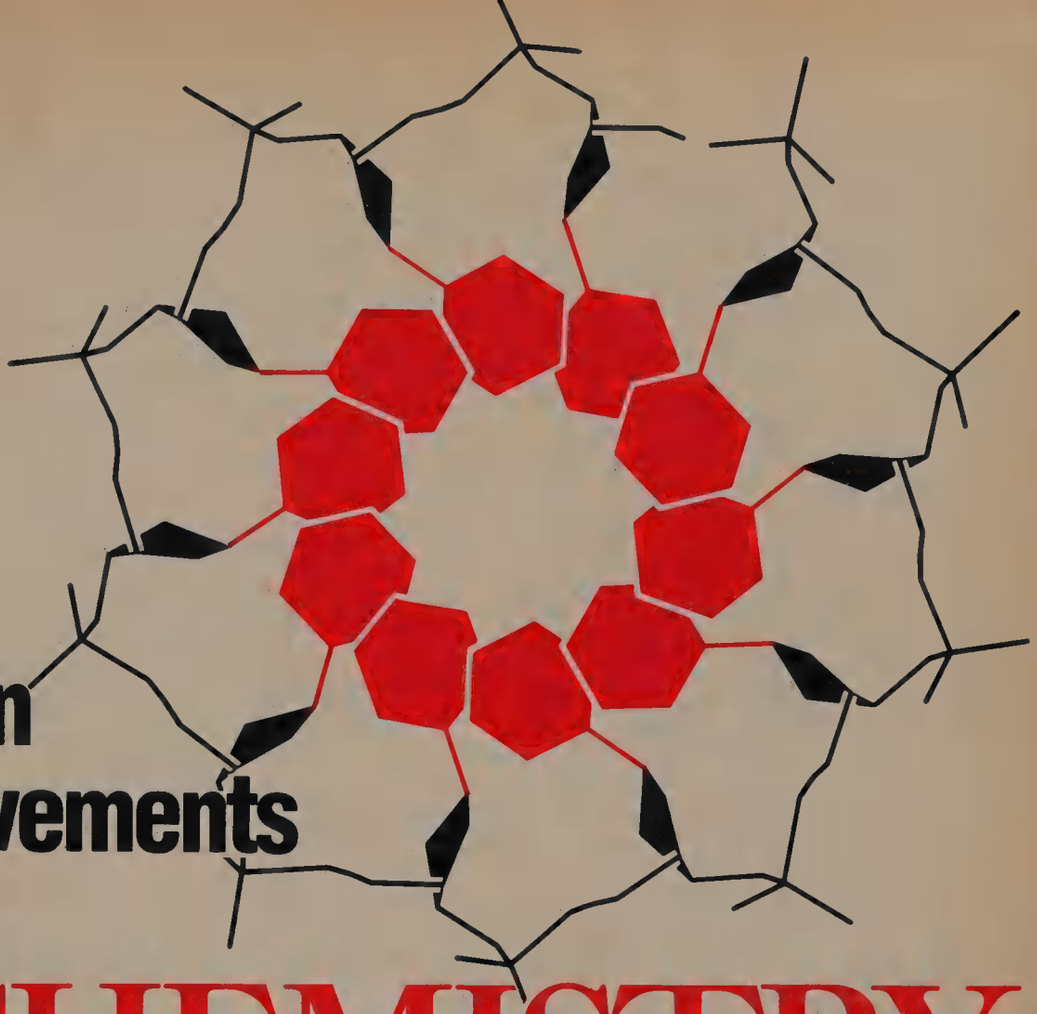


about 71,000 households in 1,100 counties to get a Current Population Report, which it can use for projections. On that basis the population of the U.S., augmented by the Armed Forces overseas, will be some 234 million, plus or minus about .8 million, on July 1, 1983. This figure is up about 1 percent per year from 1980, the net result of a birth rate of about 1.6 percent, a death rate of .85 percent and a net immigration rate of .2 percent. The projection for the year 2000 is 268 million, give or take 12 or 15 million.

To count is human, and by no means easy. The small book by the sociologist Andrew Hacker and his colleague is a critical presentation of the results of the census and of many other continuing statistical series, offered both as a book to read through and as a useful reference. Much of the material is tabular, but the tables are set into a running comment and comparison that is quite readable. The focus is demographic: how many of us there are, where we live, the vital events of birth and death, health, households, marriage and divorce. Like the census itself, however, *U/S* treats jobs and incomes, taxes and government, crime, education, housing and travel. Major questions are included in full, such as income distribution by family, by sex and race, by occupation and source, and much more.

A little cheerful numerical gossip also enters, for example the compensation of top corporate executives in 1981. Half a dozen tycoons collected more than $1 million, like top professional outfielders, even though one or two of the companies they headed were far from Easy Street. (Mobil and Union Oil could afford it.) In sum, the book offers a skeptical yet smoothed entry into what remains a rather austere assessment of American life, and it deserves both readers and browsers. It has an excellent index and a careful list of sources.

The second volume reviewed here is another matter: the annual official compendium of some 1,600 tables and quite a few graphs in 1,000 close-set pages. It was commented on in this column in the edition of 1972, after the preceding census. This time around it is still a big bargain, and its print and paper are decidedly better. It presents the unadorned tables of its sources without comment, although it carefully describes, perhaps too formally, the samples and their pitfalls. The Census Bureau undertakes to compile statistics from a wide variety of Government and non-Government sources, from the United Nations to the Federal Bureau of Investigation and from the Corporation for Public Broadcasting to *Jane's Fighting Ships*. It is plain that data are viewed as power; every bureau and industrial association tends to count the items in its own



A celebration of the achievements of a science

BIOCHEMISTRY

SECOND EDITION

LUBERT STRYER

Stanford University

949 pages
983 illustrations
$32.95

Lubert Stryer's *Biochemistry* has become the standard against which all other biochemistry books should be measured. It has shaped and continues to shape the subject it teaches.

First published eight years ago, *Biochemistry* immediately set new standards in the teaching and presentation of biochemistry. Emphasizing the process of discovery, Stryer aimed "to provide an introduction to the principles of biochemistry that gives the reader a command of its concepts and language." Reviewers commented that it was "doubtful whether there will be another text which will appeal to the student as strongly as this one does" (*Nature*).

The Second Edition, "bigger and even more splendid," is "a most handsome and thought-provoking new edition of a fine textbook" (*TIBS*). "It continues the tradition of excellence established in his first edition" (*American Scientist*), brilliantly achieving the author's aim as stated in the Preface: "to weave new knowledge into the fabric of the text...to enhance the book's teaching effectiveness by centering the exposition of new material on common themes wherever feasible and by citing recurrent motifs....and to convey a sense of the intellectual power and beauty of the discipline of biochemistry."

With a clarity and conciseness achieved through the unified narrative of a single author, "....the brevity of the text often belies its depth: the author has a unique ability to expound difficult concepts in a few words." (*Nature*)

Stryer's *Biochemistry,* Second Edition—giving up-to-date coverage of new developments while communicating the excitement and progress of biochemical frontiers.

If you are not using Stryer, compare *Biochemistry* with your present text. After the comparison, even *more* the choice.



W. H. Freeman and Company
41 Madison Avenue
New York, N.Y. 10010

universe, from mobile homes to violent assaults.

They are all here. It is the catholicity of the tables that wins the count-intoxicated reader. The catch of menhaden remains the biggest fishery in point of weight landed, as it has been for decades, although its value as a source of fish meal is only a few cents per pound. The sunflower has risen to the second most valued crop in the state of North Dakota; it is a monopoly of that state, at least as a major crop. The Federal agency with the largest budget for space was the National Aeronautics and Space Administration only from 1961 to 1981; before and since that 20-year period the Department of Defense led. In 1981 more musical works were registered for copyright than any other category of work, including serials and monographs, if the Library of Congress is right. Nearly 60 million people either hunted or fished legally for sport in 1980. Total food consumption per head remained the same (to an accuracy of 1 percent!) from 1970 to 1980. The most conspicuous changes in the diet appear to be the rise of chicken, the decline of eggs, fluid milk and cream, an increase in cheese and in fresh fruit and a surprising fall in refined sugar (more than compensated, however, by a rise in corn sweetener).

Crimes went steadily up from 1970 to a peak in 1980, with a small decrease since then. The gross death rate has been flat for 20 years and more, the rising life expectancy about keeping pace with the aging population. Newspaper circulation has been flat since 1970, and so has television viewing (except that the television has gone from black and white to color). The number of overseas telephone calls has risen by a factor of about eight. Per capita mail volume has gone up by about 15 percent over the decade; most of that rise (as the reader may have noted) is in third-class items. Union membership has grown very slowly in the past 10 years, much more slowly than the labor force. Women have found work as barbers, lawyers and computer specialists, to cite the few occupations in which there have been sizable increases in the employment of women. For minority employees office management and aircraft maintenance show a similar record. Serious and trivial together, these pages exemplify what it is fashionable to call a data base. Some base!

There is one disappointment. Last time one could pore over the economic and demographic parameters of a couple of hundred urban centers, to compare them easily. Alas, although the new list of urban centers is complete with populations, all the other comparative data are displaced into a supplementary volume: *State and Metropolitan Area Data Book*. It is not easy to characterize urban centers; they are complex. If one examines the pages carefully, it appears that great New York, with 16 million souls from New Brunswick to Montauk, is at the top of a list of 11 leading centers and Cleveland-Akron-Lorain, with just under three million, is at the bottom.

As before, the volume is not parochial; it includes a fine set of comparative international statistics, most of the economic data from UN sources. On a long list of countries Saudi Arabia leads in the relative economic importance of construction and is last in that of agriculture. Such a fact is not really surprising, but it expresses what one might surmise with an economy that is all but poetic.

MONUMENTS OF THE INCAS, text by John Hemming, photographs by Edward Ranney. A New York Graphic Society Book. Little, Brown and Company ($45). The joint effort of an American photographer and an English author, this extremely attractive volume transcends the merely showy to offer an honestly comprehensive if elegant photographic documentation of Tawantinsuyu, the empire built by the Incas within less than a century. What is left of its monuments lies beautiful and enigmatic before the lens, against the snowy peaks of the Andes along a sweeping arc from broad Lake Titicaca north to equatorial Quito. There are plenty of books about that empire conquered near its zenith, and images of its remains are commonplace, but Ranney has done more than any other photographer to show the works of the Incas as they are set into the numinous mountain landscapes that somehow called forth those superb forms in stone.

The first chapter of 15 sketches Inca history and presents an overview of Inca architecture, drawing explicitly on the admirable recent monograph of Graziano Gasparini and Luise Margolies, reviewed here (in translation) in March, 1981. Then the chapters treat, one after another, the famous localities more or less as they might be visited by a lucky traveler heading north from the Island of the Sun all the way to Ingapirca, the only surviving Inca ruin in Ecuador. About two-thirds of the book is devoted to the central region around Cuzco, that mountain capital planned in the form of a puma, here studied from Raqchi north to Vilcabamba. The illustrations are evocative and knowing, no less for a key carved detail than for a vista of meadow and peak. The text is sensitive and well documented; telling citations from the old chroniclers jostle the archaeological reports and the travelers' traditions. Hemming makes good use of the indispensable work of contemporary Peruvian scholarship; the footnotes are full and the bibliography is a find for serious beginners.

An armchair reader, envious of the fortunate who will visit the sites with this book behind them as a kind of hyperguide, found on page after page remarks and particularly scenes that escape notice in many a worthwhile book on the Inca world. The snow-covered peaks grandly tower over the rolling meadows of Soccllacasa pass, "near the rock shrines of Saihuite." There two rocks lie alone in the high open rangeland, about 100 miles toward the sea from Cuzco. The principal stone of the pair "looks from a distance like a broken flint." It lies where it has lain for a very long time, a boulder of 200 or 300 tons, never much moved by human hands. But that broken surface 15 feet across has been fashioned into "an intricate mass of carving." Small figures, pumas and hunters, lizards, frogs, snakes, vicuñas and more, have been released there, all set into a labyrinth of tiny stairs, altars and platforms, through which a network of grooves once led liquid down from the highest point in some unknown "ceremony of divination." Iconoclasts have defaced the figures, save for one brooding puma. The two photographs of this monument are extraordinary; it is a statement as timeless as it is contemporary. A microcosm was somehow carved out by the Inca's craftsmen; the 19th-century travelers who saw an Andean topographic map in that carving were not far wrong.

The vista of the real city of Machu Picchu from the southern rise of its saddle in the mountains, a familiar and beautiful shot, given us by Ranney also, seems now at a glance only one more Saihuite rock. Notches in the stone rim offer evidence that the face of the Saihuite boulder once had a copper or gilded-copper cladding, probably worked at a level of detail the relief carving in stone can never transmit to us. Perhaps the golden cities of the tall tales were real after all, even if they were only models. The strange gnomonlike platforms called inti-huatanas are called "hitching posts of the sun." Several are shown as they stand now in the Cuzco region. They are not, however, working gnomons or zenith markers as far as anyone can now make out; it seems a reasonable conjecture that here too what we now see is only a carefully carved stone base, its surfaces rounded and bland, never itself an astronomical instrument but only a kind of armature for a more intricate and controlled metal cladding. Together were they not working solar instruments?

The photographs and their excellent reproduction in duotone are so beautiful that they satisfy entirely; the work is a black-and-white prize and an archaeological introduction of real merit.



Introducing the new Osborne™ Executive™ Series.

We listened when you told us you wanted a personal business computer.

We delivered the Osborne 1.

And in the two years since its introduction, the Osborne 1 has become the largest selling portable business computer on earth.

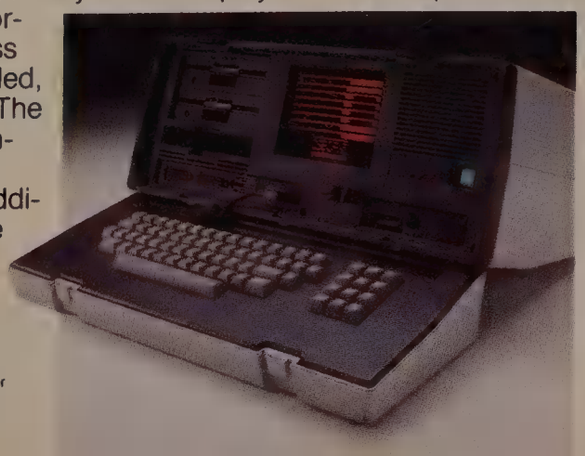
We listened harder.

The result is the new Osborne Executive Series. Consider these new advantages:

☐ A big, 80-column, eye-easy amber display lets you see more of the information you're using, with less effort. ☐ Memory is expanded, for larger spreadsheets. ☐ The new Executive Series now includes a database system—Personal Pearl.™ (That's in addition to an already formidable list of standard Osborne software, including word processing and electronic spreadsheets.) ☐ Terminal Emulation provides access to information stored on other computers. ☐ IBM Personal Computer™ compatibility (on the Executive 2) lets you use software packages designed for the IBM system.

And there's much more.

Add up the value of all the hardware features and software systems included with either the Executive or Executive 2 and you will see that the sticker price is one of the incredible bargains available in personal computing.



Visit your local authorized Osborne dealer. Once again, Osborne has set the standard for portable business computers.

For the name of your nearest Osborne dealer, dial (800) 772-3545 ext. 905 (in California); (800) 227-1617 ext. 905 (outside California).

OSBORNE™
COMPUTER CORPORATION

Trademarks: Osborne 1, Executive, Executive 2: Osborne Computer Corporation. Personal Pearl: Relational Software. IBM Personal Computer: International Business Machines.



THE WORLD'S FASTEST BAGGAGE HANDLER?

If there's a station wagon in the world that's faster than the Turbo Wagon from Volvo, it's a very well-kept secret.

The Turbo Wagon can take you from a standing start to the legal speed limit in 9.2 seconds. As *Motor Trend* puts it: "You keep hoping for a stray Saab 900 Turbo or even a BMW you can sniff out and then send scrambling to the roadside while you blow its doors in."

We assure you, a test drive in the Turbo Wagon will be very revealing.

Especially if you thought the only time a station wagon could haul was when it had something in the back.

THE TURBO WAGON

By Volvo.

Established 1845 SCIENTIFIC AMERICAN July 1983 Volume 249 Number 1

The Slowing of Urbanization in the U.S.

The rural population is growing as fast as the urban population for the first time. The geographic expansion and coalescence of urban areas, however, has led to the appearance of massive "supercities"

by Larry Long and Diana DeAre

One of the most fundamental features of a modern society is the demographic predominance of the cities over the countryside. Urbanization, along with the industrial growth that has generally accompanied it, has historically distinguished the developed countries from the developing ones. In a country that is undergoing economic development two tendencies usually contribute to the increasing dominance of the urban areas. First, the population of established cities grows faster than the population of the nation as a whole. Second, in some rural areas population growth is rapid enough for the areas to become urban. In the U.S. both of these basic forms of urbanization have been operating almost continuously at least since the decennial census was initiated in 1790.

The detailed results of the census of 1980, however, show that urbanization in this traditional twofold form was strongly modified in the U.S. between 1970 and 1980. The first form of urbanization ceased during the decade: the population of rural areas grew as fast as that of urban areas. Only the second kind of urbanization is still in progress: between 1970 and 1980 the population of many rural areas grew enough for them to be reclassified as urban. If it had not been for growth in areas that were considered rural in 1970, the fraction of the U.S. population living in urban areas would have decreased for the first time. As the growth of cities was largely attributable to economic factors, so too the decline appears to have economic causes. The available data suggest that the rural areas' attractiveness for industry is responsible for much of their population growth.

Although the deceleration of urban growth is widespread, it is by no means geographically uniform. In the Northeast both kinds of urbanization have ended. In the South and West both kinds are still in progress, but the overall rate of urbanization is lower than it was in the past. Moreover, the current trends have quite different effects on the various geographic components of a city. In the 1950's and 1960's a common pattern was for the city proper to lose population while the suburbs gained. In many instances the city itself continued to lose people in the 1970's and the area subject to population loss expanded to include the inner suburbs. The outer suburbs, however, continued to grow both in population and in area. As a result the urban areas of the U.S. have become larger and less densely populated.

Urban areas that are expanding in area can ultimately run into one another. The coalescence of neighboring cities in the 1970's led to an increase in the number of "supercities": continuously built-up urban areas with a population of a million or more. The growth of the supercities shows that the current trends do not necessarily mean the end of large urban centers. Recent developments do suggest, however, that the economic advantages of locating industry in cities may have been reduced. As a result new and less concentrated forms of human settlement may be in prospect.

Measuring the number of people who live in urban areas requires a definition of what is urban, which would appear to be a straightforward matter. One could, for example, simply count the people who live within the boundaries of cities of a certain size. Such a definition, however, would omit many of the residents of the suburbs, who are certainly part of the urban population. Moreover, the question would naturally arise of how many residents a city must have before it is included in the count.

Because of such ambiguities there are many possible methods of distinguishing the urban population from the rural. The Bureau of the Census has several ways of identifying the urban part of the U.S. The oldest definition includes all incorporated places with a population of 2,500 or more. Such a broad category almost inevitably includes some areas that do not correspond to the ordinary notion of what is urban.

The Census Bureau also employs a narrower category that more closely approximates the usual notion. The second category defines an urbanized area as a city with a population of 50,000 or more together with all the contiguous built-up land that has a population density of at least 1,000 per square mile. The large city included in the urbanized area is termed the central city. In this article we refer to the suburbs that share a boundary line with the city as the inner suburbs; all other parts of the urbanized area are referred to as the outer suburbs.

The concept of the urbanized area was introduced in analyzing the results of the census of 1950. The motivation for its introduction was the rapid growth of the suburbs in the years after World War II. The urbanized area provides a

good measure of the demographic limits of a city as opposed to its administrative boundaries, which often do not coincide with the extent of urbanization.

If the urbanized areas are removed from the broad category of all urban places, what remains is the group of nonsurburban cities and towns with a population of more than 2,500 but less than 50,000. In the terminology of the Census Bureau such areas are referred to as "other urban" places; here we shall call them the small cities and towns. Territory that is outside both the urbanized areas and the small cities and towns is considered rural.

We have employed the returns from the census of 1980 to analyze recent trends in the growth of the urban and rural populations. We have examined developments in both the urbanized areas and the small cities and towns, and we shall describe some of the trends in the territory that is urban in the broad sense. We shall concentrate, however, on the urbanized areas. The reason is that included among the small cities and towns are places that in an important sense are more rural than urban. Furthermore, many small cities and towns outgrow that classification to become urbanized areas and others are absorbed in the expansion of the large cities.

The definitions of what is urban employed by the Census Bureau have undergone technical changes since 1970. We have adjusted the data from the census of 1980 to make them comparable with the data from the three preceding censuses. The adjustment has enabled us to make consistent comparisons of trends in population, geographic area and population density in the U.S. from 1950 to 1980.

The most general and significant conclusion that emerges from our work is that the level of urbanization stopped increasing for the first time in the 1970's. The population of the urbanized areas is still growing, but at a rate lower than that of the rural areas. As a result the fraction of the nation's population in urbanized areas is no longer increasing. In 1950, 46 percent of the U.S. population lived in an urbanized area; in 1960, 53 percent did so; in 1970, 58 percent. In 1980 the fraction was still 58 percent. In some parts of the country the fraction is decreasing.

If the urbanized areas are classified according to when they became large enough to be designated as such, the deceleration of the urbanizing process appears even more dramatic. The oldest urbanized areas are the ones designated in 1950. In the 1950's and 1960's these older cities grew substantially faster than the population of the nation. By the 1970's, however, the older urbanized areas, with an average annual population growth rate of .8 percent, were growing slower than the national population, which had a rate of 1.1 percent. If it had not been for the faster-than-average growth of the newer urbanized areas, the fraction of the population in the large cities would have declined.

These data imply that one of the two fundamental forms of urbanization—the growth of preexisting urban cen-



THERMAL ISLAND is shown on a map of the area around Los Angeles and Long Beach in California. The map indicates the daytime surface temperatures. The data were collected by a satellite operated by the National Aeronautics and Space Administration and were plotted by a computer. White corresponds to the highest temperature; dark green, light green and red correspond to intermediate temperatures; blue corresponds to the lowest temperature. A thermal island is an area of elevated temperature resulting from the high density of manmade structures. The hottest part of the region is an industrial park (*white area at upper right*). Los Angeles and Long Beach make up an urbanized area: a city with 50,000 people or more and all the contiguous territory with a density of at least 1,000 per square mile.

ters—stopped in the U.S. in the 1970's. The slowing of the growth of the older cities was very rapid. In the 1950's the rate of population growth in the older urbanized areas and those added in 1960 was 23 times the rate of areas outside them. In the 1960's the rate of growth in the older urbanized areas and those that had been added by 1970 was 16 times the rate of the rest of the country. In the 1970's the rate of growth in the urbanized areas was only marginally greater than the rate in other places, and it was higher only because of rapid growth in areas that had been considered rural in 1970. This development is unprecedented in U.S. demographic history.

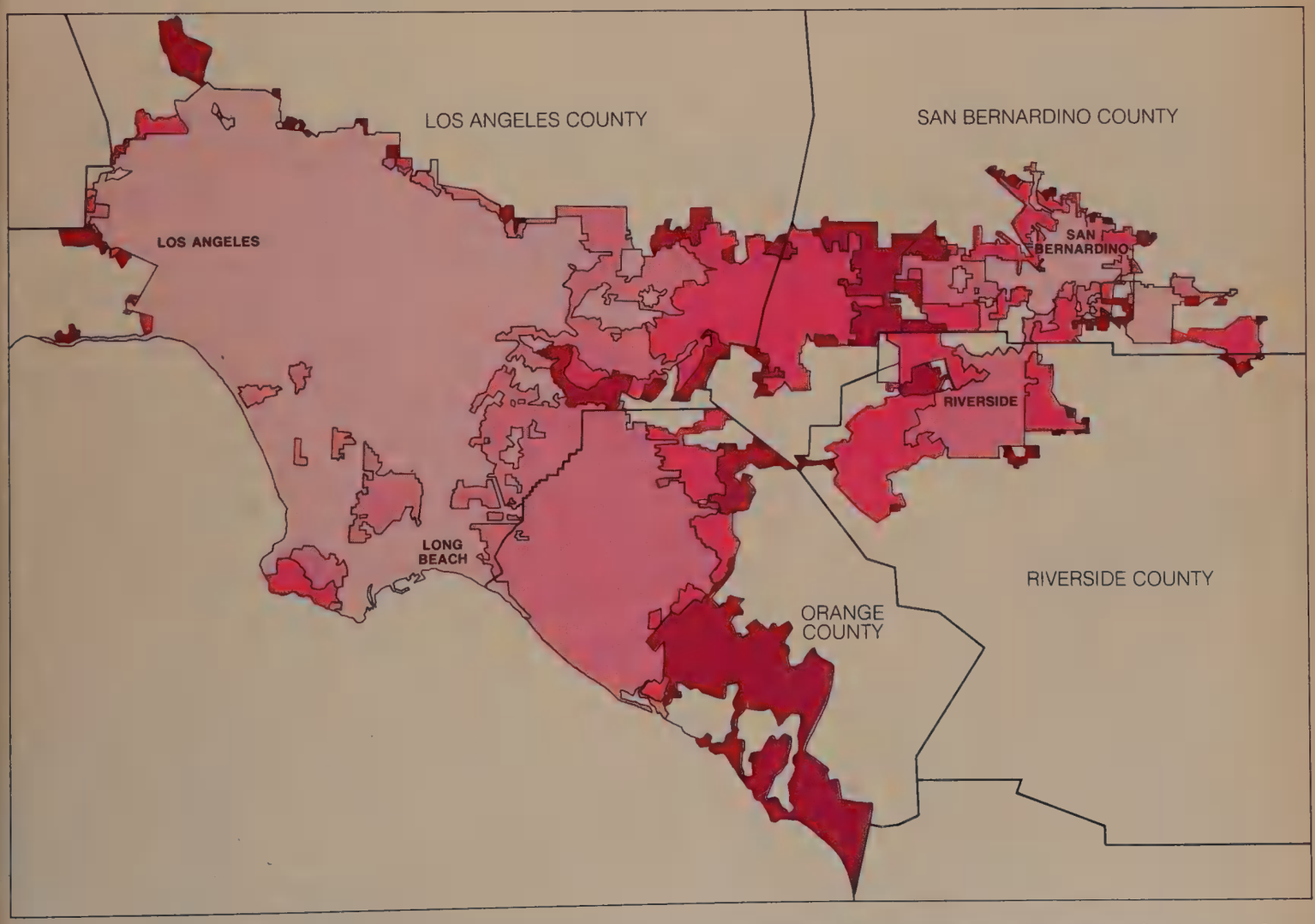
If the broad concept of what is urban is employed, the level of urbanization did not merely stop increasing in the 1970's but actually declined. If those who live in small cities and towns are included, the fraction of the U.S. population in urban areas was 64 percent in 1950. It increased to 70 percent in 1960 and to 73.5 percent in 1970 before decreasing to 73.3 percent in 1980.

A decline of less than one percentage point is not substantial; its significance is not in its size but in the fact that it comes after a long period of uninterrupted growth. An advantage of the more inclusive definition of urban territory is that it can be applied to census records extending far into the past. The level of urbanization in the broad sense declined only once before: between 1810 and 1820. That decline was the result of population growth in several rural states of the Middle West that had been added by the Louisiana Purchase.

The slowing of the growth of the urban population may be a temporary phenomenon that could be reversed by changes in economic factors such as the price of fuel. It is more probable, however, that we are witnessing a basic demographic reversal. The immediate cause of the reversal is not a decrease in the absolute size of the urban population but a rapid increase in the rate of population growth in rural areas. The rural population declined slightly in the 1950's and 1960's. In the 1970's, however, the rural population increased by 12 percent, a rate of growth about equal to that of the urbanized areas and greater than that of the small cities and towns, which increased by 9 percent. Thus the demographic growth outside the large cities took place in the open countryside rather than in the small towns.

The rate of urbanization decreased in the U.S. as a whole in the 1970's. The national trends, however, conceal significant differences among the regions of the country. For the purposes of the census the U.S. is divided into four major regions. The Northeast region includes the New England states as well as New Jersey, New York and Pennsylvania. The North Central region includes the Great Lakes states and Iowa, Kansas, Missouri, Nebraska, North Dakota and South Dakota. The South and West regions consist of the remaining states, with Texas being the westernmost state in the South region.

In the Northeast in the 1970's there



SOUTHERN CALIFORNIA SUPERCITY is made up of the Los Angeles, Long Beach, San Bernardino and Riverside urbanized areas. As the urbanized areas expanded geographically in the 1970's they ultimately encountered the boundaries of other urban places. When this happened, they merged into supercities: continuously built-up urban masses with a population of a million or more. The map shows the boundaries of the urbanized areas of southern California as determined by the censuses of 1950 (*lightest color*), 1960 (*light color*), 1970 (*dark color*) and 1980 (*darkest color*). By 1980 the Los Angeles–Long Beach area and the San Bernardino–Riverside area had merged into a massive supercity. Such expansion has generally been accompanied by a decrease in the population density of the urbanized areas.

was a decrease in the population of the urbanized areas from 34.8 million to 34.4 million. The fraction of the total population of the region in urbanized areas declined from 71 to 70 percent. Furthermore, no new urbanized areas came into being in the region during the decade. The Northeast is therefore a highly urbanized, demographically mature region that has become less urbanized because both basic forms of urbanization have ended.

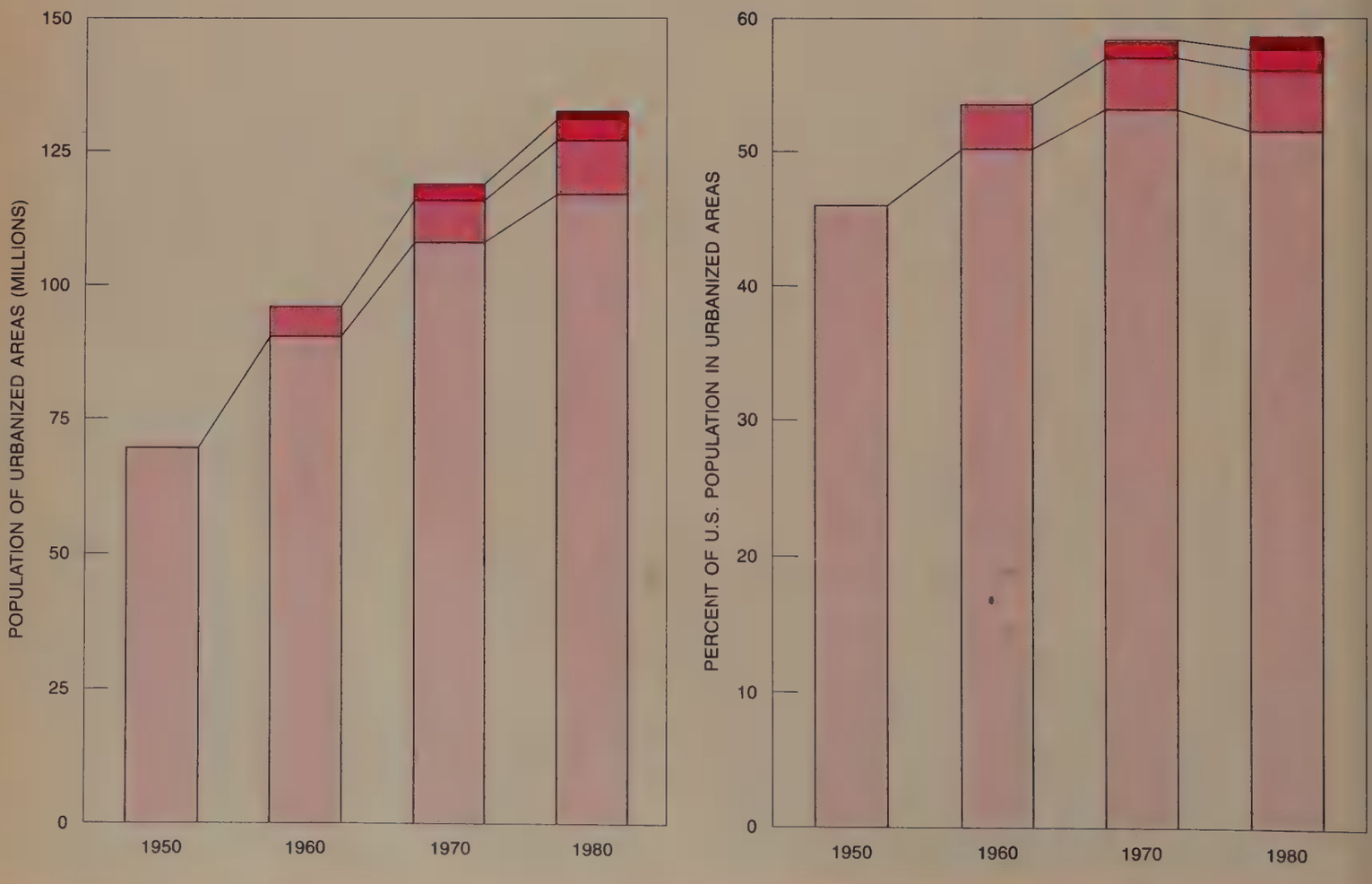
In the North Central region there was a slight increase in the population of the urbanized areas: from 31.5 to 31.9 million. The population outside the urbanized areas, however, grew at a much higher rate and as a result the fraction of the population in the urbanized areas declined from 56 to 54 percent. The newer urban centers in the North Central region (those designated in 1960 or 1970), on the other hand, grew faster than the regional average. In addition a number of places grew rapidly enough to be newly classified as urbanized areas. Hence in the North Central region the decrease in the level of urbanization was solely the result of population decreases in the older urban centers.

In the South and West the degree of urbanization increased according to the twofold traditional pattern. Intriguingly, although the West is often thought of as a region of "wide open spaces," it is in reality almost as urban as the Northeast, which is much more readily imagined as a highly urban region. Between 1970 and 1980 the fraction of the population of the West in an urbanized area increased to 69 percent; as we have indicated, the fraction in the Northeast decreased to 70 percent. This startling comparison emphasizes the difference between the demographic and the geographic measures of urbanization, a distinction that is quite important in analyzing current trends in urbanization. Although most of the land in the West is not devoted to urban uses, most of the region's inhabitants live in urban places.

It is probable that the level of urbanization in the West will for some time remain higher than the level in the North Central and South regions. In the South the population of urbanized areas increased from 29 to 36 million; in 1980 this represented 48 percent of the total population. In both the West and the South the relative increase in the population of the urbanized areas was smaller in the 1970's than it had been in the two preceding decades.

When we examined census results for individual states, we found that "ruralization," the increase in the fraction of the population in rural areas, was widespread and that it occurred in states with quite different levels of urban concentration. Ruralization occurred in 18 states in the 1970's, a substantial increase over the 1960's, when it occurred in only four. The 18 states include some where the population is growing rapidly, such as Texas and Oklahoma, and some where the total population is decreasing, such as New York. The process could be detected in both highly urban and highly rural states. Indeed, the reversal was so complete that there were only three states where the absolute size of the rural population decreased. Twenty-nine states had such losses in the 1960's, and most of the 29 had also lost rural population in the 1950's.

A vivid example of the pervasiveness



LEVEL OF URBANIZATION, or the fraction of the population in the urbanized areas, has almost stopped increasing. If the population of a small town grows enough, the town becomes an urbanized area. The subdivisions of the bars correspond to urbanized areas added after the censuses of 1950, 1960, 1970 and 1980. Between 1970 and 1980 the total population of the urbanized areas continued to grow (*left*). Because of renewed population growth in the rural areas, however, the fraction of the population in the urbanized areas increased little (*right*). The fraction would have decreased except for rapid growth in the urbanized areas added in 1980.

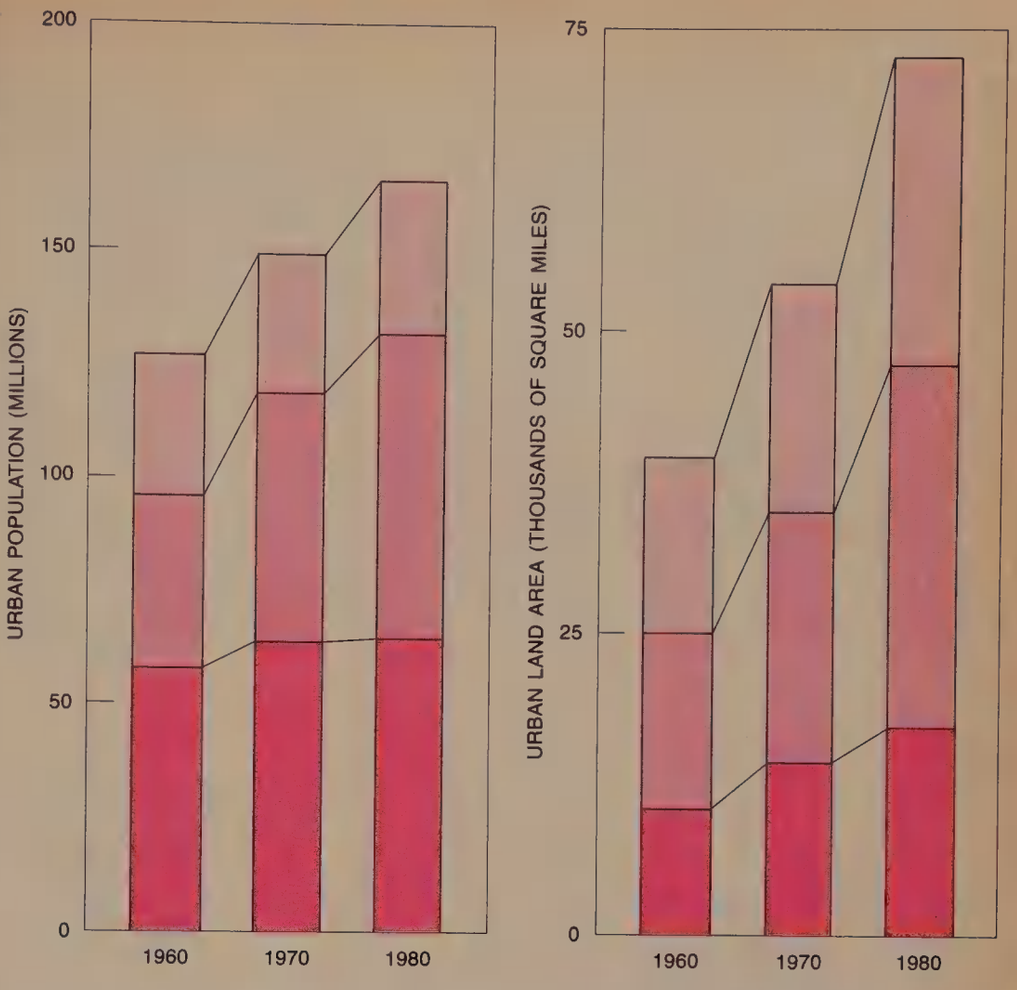
of rural demographic growth in the 1970's is provided by the section of the Eastern Seaboard often referred to as Megalopolis. The term Megalopolis was coined by Jean Gottmann of the University of Oxford in a book of the same name published in 1961. Gottmann defined the region as extending from southern New Hampshire to northern Virginia and including Boston, New York, Philadelphia, Baltimore, Washington and the smaller metropolitan areas in between. The area Gottmann specified also includes the rural areas on the fringes of the urban centers.

In 1960, 37 million people, a fifth of the nation's population, lived within the boundaries of Megalopolis. Surprisingly, between 1970 and 1980 Megalopolis stopped becoming more urban. Its rural population grew 10 times as fast as its urban population and the urban population decreased from 87 to 86 percent of the total. Demographic growth on the fringe of the urban settlements continued, but the central cities lost population. Population growth in the rural areas accelerated moderately, enough for it to surpass the overall urban rate.

It should not be assumed that the renewed demographic vitality of rural areas is due to an increase in the number of people who live and work on farms. On the contrary, the number of people who live on farms continued its long-standing decline in the 1970's. Unlike the count of population, data for employment are not generally tabulated according to urban and rural categories. The Census Bureau does, however, compile annual data on the number of jobs in each county. The information is collected from the reports employers file with the Social Security Administration and the Internal Revenue Service.

We classified each county according to whether it had an urban settlement in it. A county was considered rural if it included no settlement with a population of 2,500 or more. A county was considered metropolitan if it included a city with a population of 50,000 or part of the suburbs of such a city. Between the two extreme categories of metropolitan and rural counties are many counties whose largest settlement is one of the "other urban" areas. When the counties had been classified, we counted the new jobs brought into being in each one by the renewed economic activity that followed the recession of 1974–75.

Our analysis suggests that one of the major factors in the deconcentration of the U.S. population is the fact that new jobs are being added as rapidly in rural areas as they are in urban areas. Overall the rate of increase of employment was greater in the rural counties. Furthermore, the rural rate was greater in many kinds of businesses, including manufacturing companies. In some regions and in some industries the rate of growth in jobs was far greater in the rural counties than it was in the metropolitan counties.



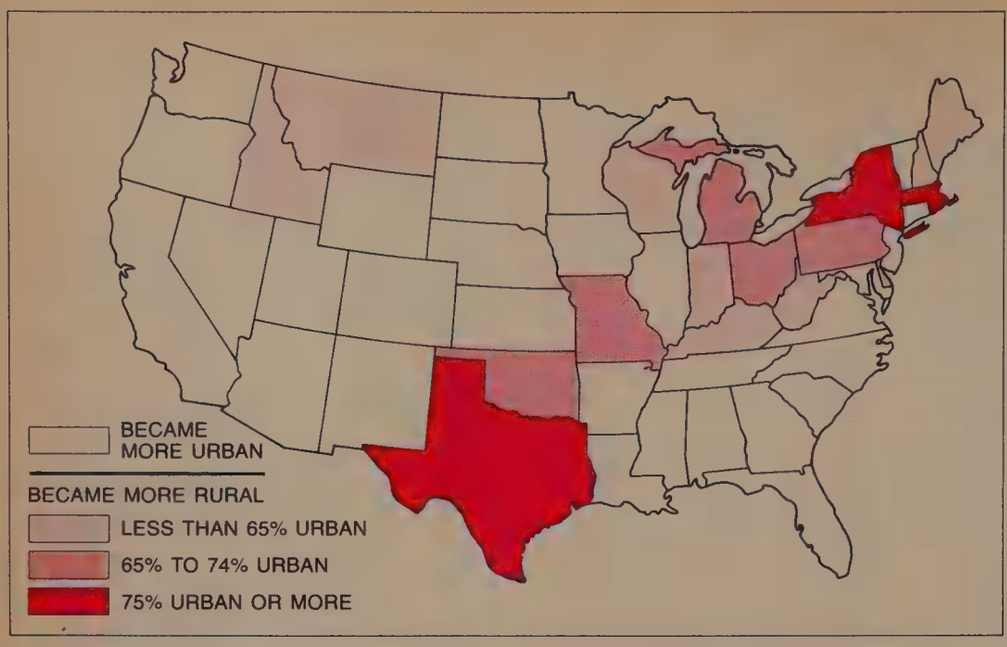
GEOGRAPHIC COMPONENTS of urban territory have been affected quite differently by recent population trends. Central cities (*dark color*) are the major cities included in the urbanized areas. The suburban fringe (*intermediate color*) makes up the remainder of an urbanized area. Small cities and towns (*light color*) are those with a population of from 2,500 to 50,000. From 1960 to 1980 the total urban population increased by 33 percent (*left*). The total urban land area increased by 84 percent in the 20-year period (*right*). The result was a sharp decrease in the population density of urban territory. The population and land area of the suburbs increased as a proportion of the total. The population and land area of the other types of urban territory (the central cities and the small cities and towns) decreased as a proportion of the total.

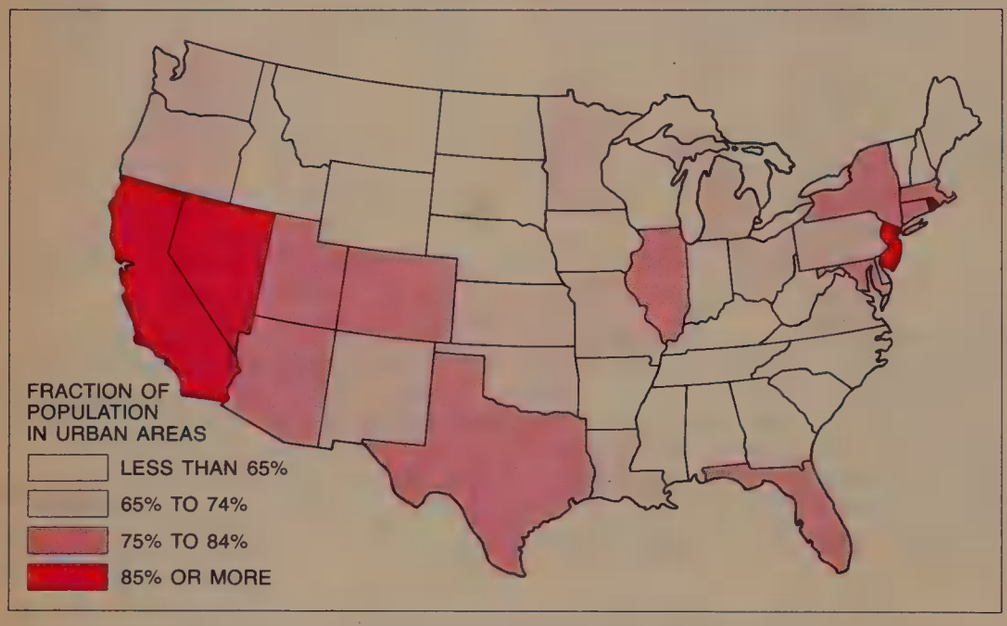
Thus economic forces could be responsible for much of the deceleration of the growth of the urban population. Although population growth in the urbanized areas has slowed, the areas have continued to expand rapidly in a geographic sense. The disjunction between demographic and geographic growth has led to significant changes in patterns of land use and population density.

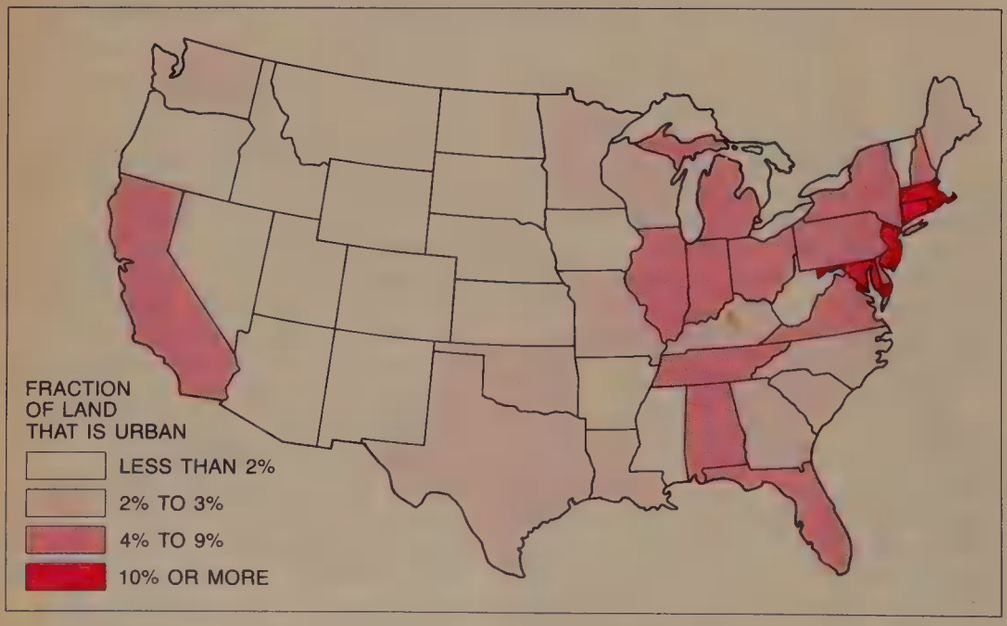
Employing the broad definition of what is urban, about 14,400 square miles were converted from rural to urban uses in the 1960's; in the 1970's about 18,700 square miles were converted. Some 74,000 square miles, or 2 percent of all land in the U.S., is now covered by cities, suburbs and small towns. It is readily apparent that the increase in urban land area has been much faster than the increase in urban population. From 1960 to 1980 the area of urban settlement increased by 84 percent while the urban population increased by only 33 percent. What is more, the disparity between the demographic and the geographic rates is increasing. In the 1960's the rate of spatial expansion was not quite twice the rate of population growth; in the 1970's the spatial rate was three times the demographic one.

If the area of urban settlement is increasing faster than the urban population, it follows that the density of the urban population is decreasing. Before turning to the effect of such a trend on the components of the urbanized areas, we shall briefly consider the relation between population and geographic area at the state level. Analyzing urban growth in terms of both land use and population can provide a necessary corrective to an exclusively demographic approach. For example, the state with the greatest fraction of its population in urban areas is California. The West is highly urbanized, but California is exceptional even in the region: 91 percent of its population is in urban areas, broadly defined. Geographically, however, California is rural: 96 percent of its land is outside the urban areas.

New Jersey presents a contrasting picture. The fraction of the state's population that is urban is about the same as it







RURALIZATION, or an increase in the fraction of the population in rural areas, occurred in 18 states in the 1970's (*top*). This was a substantial increase from the 1960's, when only four states became more rural. The 18 were in all four census regions; they included both highly urban and highly rural states. Geographic and demographic measures of urbanization yield quite different pictures of the states. The demographic measure is the fraction of the state population in cities and towns (*middle*). By this measure California is the most urban state. The geographic measure is the fraction of the state's land area covered by cities or towns (*bottom*). By this measure California is much less urban than New Jersey, the most urban of the states.

is in California (89 percent) but a smaller fraction of land is rural (71 percent). Hence New Jersey is the most densely populated state. Its residential density of 986 people per square mile is greater than that of Belgium, one of the most densely populated nations in Europe.

When one considers a typical New Jersey landscape, it might seem that much more than 29 percent of the state's land is urban. The definition of what is urban, however, refers to residential density and therefore excludes certain areas that might be counted as urban by other standards, such as land given over to factories, oil refineries, storage facilities, industrial parks and highways that lie outside urbanized areas and the corporate boundaries of towns or cities with more than 2,500 inhabitants. The land under such facilities is rarely considered rural, but it is not urban in the demographer's sense because it provides homes for very few people.

As we observed at the beginning, urban growth and industrial development have historically been closely connected. In the traditional pattern of industrial development factories were built in or at least near urban centers. The factors that affect where industrial plants are constructed have apparently been changing in recent years. The result has been the rapid growth of industry in rural settings. The people who moved to the rural areas to take up the new jobs there, however, do not conform to the ordinary image of the rural population. Rather than being farm workers, they are in many instances either workers with the kind of technical training required in modern industry or professionals. The new rural demographic concentrations appear to represent small centers of urban culture transplanted to the countryside and enabled to survive by recent advances in communications, transportation and methods of industrial production.

There has been a substantial decline in the average population density of the urban areas. The density of urban places in the broad sense decreased from 3,200 people per square mile in 1960 to 2,300 in 1980. The decline was faster in the 1970's than it was in the 1960's. Since 1950 the average density of urbanized areas has been reduced by almost half.

The reduction in density has two causes. First, existing urban places have become less densely populated. In addition new urban areas generally have a lower density than older ones. The reduction in urban population density together with the growth of the rural population has led to a slight narrowing of the differential in density between the urban and the rural populations. The rural population, however, with a density of 17 people per square mile, is still much sparser than the urban population.

Declines in urban population density

were widespread in both the 1960's and the 1970's. In the 1960's 81 percent of all urbanized areas became less densely populated; in the 1970's the fraction was 87 percent. The exceptions were for the most part cities where geographic circumstances prevented further expansion. For example, the population density of Honolulu increased because territorial growth is made difficult by mountains and the seacoast. The density of the urbanized area that includes Miami, Fort Lauderdale and West Palm Beach increased because the three communities have merged and population increases must be accommodated by the filling in of vacant suburban land.

The coalescence of the three Florida cities has resulted in the formation of an urban mass with a population of more than three million covering more than 800 square miles. The formation of this mass exemplifies a process that continued in the 1970's. As the urbanized areas have expanded geographically they have in many instances reached the boundaries of adjacent cities. The contiguous urban areas then join to form amorphous urban regions covering large amounts of land.

Unlike Megalopolis, which includes a substantial rural periphery, these supercities represent continuously built-up urban territory. There are now 29 urban centers in the U.S. with a population of a million or more. About half of them were created by the fusion of two or more preexisting cities. The supercities are not recognized as legal or political entities, although they are interdependent in planning for the provision of such services as public transportation, sewage and water.

The supercities represent a demographically vital part of urban territory. Even as the rate of urbanization decreased, the proportion of the population living in the supercities increased. In 1970 there were 24 supercities with a population of 76 million, or 37 percent of the U.S. population. In 1980 the population of the 29 supercities was 89 million, or 39 percent of the total. The supercities have expanded by incremental growth at their suburban fringes, by

SUPERCITIES, the urban masses with a population of a million or more, include both incorporated areas and built-up suburban territory. In many cases they have been brought into being by the merging of two or more large cities. In 1970 there were 24 supercities; by 1980 there were the 29 shown here. The development of the supercities shows that the peaking of urbanization need not halt the development of very large urban centers. All the supercities increased in area during the 1970's, but several lost population (*bold type*). Although some cities within the boundaries of large urban complexes have been losing population for decades, not until the 1970's did any of the large merged urban masses lose residents.

SUPERCITY	1980 POPULATION	LAND AREA (SQUARE MILES)	POPULATION DENSITY (PEOPLE PER SQUARE MILE)
NEW YORK CITY/ NORTHEASTERN NEW JERSEY/ SOUTHERN CONNECTICUT	17,606,680	3,656	4,816
LOS ANGELES/LONG BEACH/ SAN BERNARDINO/RIVERSIDE	10,184,611	2,187	4,657
CHICAGO/NORTHWEST INDIANA/ AURORA/ELGIN/JOLIET	7,212,778	1,675	4,306
PHILADELPHIA/ WILMINGTON/TRENTON	4,779,796	1,270	3,764
SAN FRANCISCO/SAN JOSE	4,434,650	1,122	3,952
DETROIT	3,809,327	1,044	3,649
BOSTON/BROCKTON/LOWELL/ LAWRENCE/HAVERHILL	3,225,386	1,084	2,975
MIAMI/FORT LAUDERDALE/ HOLLYWOOD/WEST PALM BEACH	3,103,729	816	3,804
WASHINGTON, D.C.	2,763,105	807	3,424
HOUSTON/ TEXAS CITY/LAMARQUE	2,521,857	1,169	2,157
CLEVELAND/AKRON/ LORAIN/ELYRIA	2,493,475	989	2,521
DALLAS/FORT WORTH	2,451,390	1,280	1,915
ST. LOUIS	1,848,590	597	3,096
PITTSBURGH	1,810,038	713	2,539
SEATTLE/ EVERETT/TACOMA	1,793,612	672	2,669
MINNEAPOLIS/ST. PAUL	1,787,564	980	1,824
BALTIMORE	1,755,477	523	3,357
SAN DIEGO	1,704,352	611	2,789
ATLANTA	1,613,357	905	1,783
PHOENIX	1,409,279	641	2,199
TAMPA/ST. PETERSBURG	1,354,249	527	2,570
DENVER	1,352,070	439	3,080
CINCINNATI/HAMILTON	1,228,438	465	2,642
MILWAUKEE	1,207,008	496	2,433
NEWPORT NEWS/HAMPTON/ NORFOLK/PORTSMOUTH	1,099,360	614	1,790
KANSAS CITY	1,097,793	589	1,864
NEW ORLEANS	1,078,299	230	4,688
PORTLAND	1,026,144	349	2,940
BUFFALO	1,002,285	266	3,768

absorbing smaller cities and by fusion with cities of equal size. The process by which such places have developed indicates that the slowing of urbanization need not entail the disappearance of large urban centers.

Examining the continuously built-up urbanized areas helps to overcome the difficulties inherent in taking as the unit of analysis the city proper, cities whose boundaries can be changed only with difficulty. The very large urbanized area in New York, northern New Jersey and southern Connecticut that includes New York City is still the nation's largest supercity. Its population in 1980 was 17.6 million, and although it lost people during the decade, it continued to expand geographically. The importance of the distinction between a city and an urbanized area can be shown by comparing Chicago with Los Angeles. Chicago itself has more residents than the city of Los Angeles, but the urbanized area around Los Angeles surpassed the one around Chicago in population during the 1950's. By 1980 urban development had joined Los Angeles, Long Beach, San Bernardino, Riverside and many smaller places. The result is an area that is geographically as well as demographically larger than the urban nexus that includes Chicago, Aurora, Elgin and Joliet in Illinois and Gary in Indiana.

One tends to think of southern California as a more or less suburban area with a low population density. The Los Angeles supercity, however, has a population of about 10 million on 2,000 square miles of land. Its density of 4,700 people per square mile is greater than the densities of the supercities around Chicago, Philadelphia, Boston and several other older urban areas.

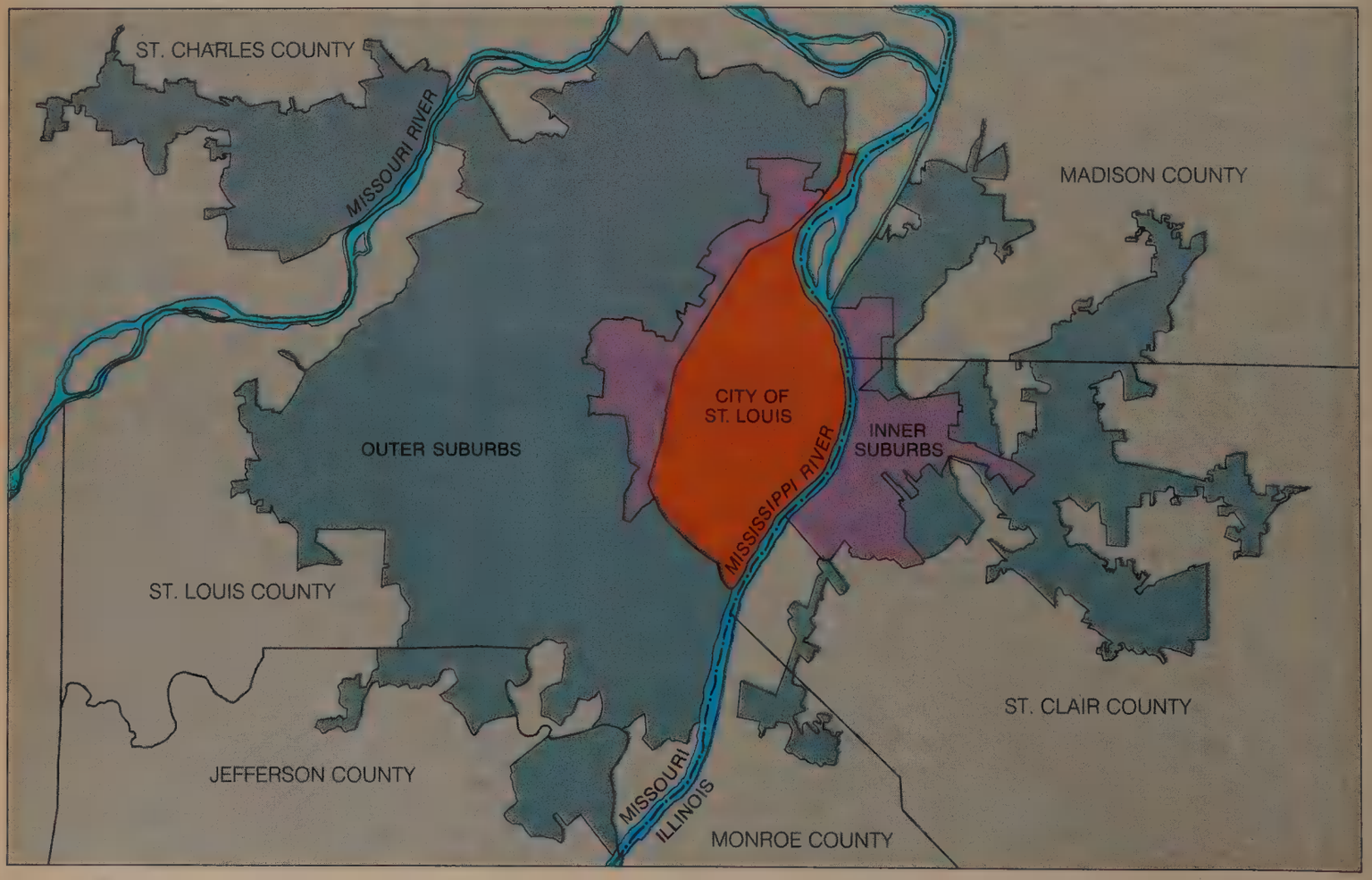
Density is high in California because the population is growing and spatial expansion is limited by the coastline, the mountains and the large areas that are already taken up by dense settlement and highways. An observer may be less conscious of the density of population in southern California than he is of that in the other regions because the steep gradients of population density that distinguish cities in the Northeast or the Middle West from their suburbs do not exist on the West Coast. Nevertheless, the urbanized part of southern California is more densely populated on the average than many major urban centers in the Northeast and the Middle West are.

The changes in density and geographic area in the 1970's have affected all parts of the urbanized areas from the central cities to the outer suburbs. The changes, however, have not affected the components in the same way: a disproportionate amount of growth in both population and area has taken place in the suburbs. In 1960, 30 percent of the total urban population lived in the suburbs, which occupied 36 percent of urban land. By 1980, 41 percent of urban residents lived in the suburbs, which occupied 42 percent of urban land. In the same period the fraction of the urban population living in the central cities decreased, and so did the fraction living in small cities and towns.

The increase in the suburban share of the urban population has been even more rapid in the older cities than the national data suggest. In 1950 about 30 percent of the population of the urbanized areas lived in the suburbs; in 1980 more than 50 percent did. The increase has been slower in the newer urbanized areas, partly because in many instances a new city can annex surrounding territory. Many older cities are surrounded by incorporated areas that can resist annexation; hence geographic expansion of the central city is limited.

Because the suburbs were areas of relatively rapid demographic growth in the 1970's, their population density decreased less than that of other places. The average density of the suburbs decreased by 15 percent, that of the central cities by 17 percent and that of the small cities and towns by 18 percent. The changes were too small to affect the overall ratio of population density between the cities and the suburbs, which was 1.7 in both 1970 and 1980.

The fact that the ratio of population density in the cities and suburbs did not change conceals some significant demographic trends. In many urbanized areas the central city has been losing popula-



tion for decades. In general the urbanized area as a whole has gained population because the growth in the suburbs has compensated for the losses in the central city. The 1980 census has shown, however, that in many cases an entire urbanized area is now losing population. One reason is that the rate of population loss in the central city has increased. A second reason is that the suburbs closest to the central city have begun losing population.

The urbanized area that includes St. Louis provides a good example of the process. The city of St. Louis lost population in both the 1960's and the 1970's; the rate of loss was higher in the 1970's. The losses have been so great and so sustained that the city now has roughly the same number of residents it had in 1890. Like most other cities in the Middle West and Northeast, St. Louis has long been surrounded by incorporated areas. These inner suburbs began to lose people in the 1960's, and their population declined by about 15 percent in the 1970's. Beyond the ring of inner suburbs population is still growing, but at a sharply diminishing rate. St. Louis' outer suburbs grew by 103 percent in the 1950's, by 55 percent in the 1960's and by 17 percent in the 1970's. Other Middle Western and Northeastern cities show a similar pattern.

The image of a doughnut is often employed to describe an urbanized area with an older central city whose population is decreasing. The image is not new; such population losses have been going on for decades. What is new is the expansion of the region of population loss (the "hole" in the doughnut) to include the inner suburbs. The hole has grown largest in the urbanized areas where the doughnut pattern has prevailed longest.

Moreover, although the doughnut has long been a familiar image in the industrial centers of the Northeast, in the 1970's it spread for the first time to much of the rest of the U.S. About half of all urbanized areas show the pattern, including a fourth of the cities in the South. The city of Atlanta, after growing in the 1950's and 1960's, lost population in the 1970's at a rate greater than that for Newark, N.J.

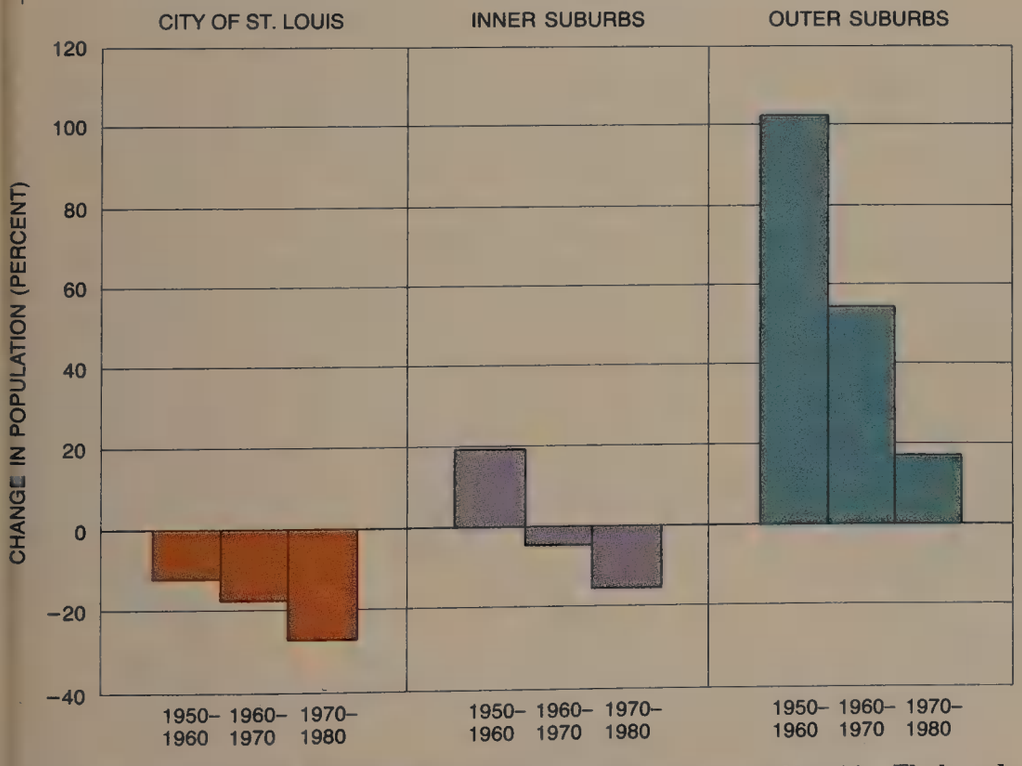
Population losses in the central cities and inner suburbs have become so great that an increasing number of urbanized areas are losing population. In the 1950's only one area in 20 lost residents; in the 1970's one in six did. Most such urban population declines are invisible to the casual observer. Indeed, in the light of "gentrification" (the movement of young professionals into inner-city neighborhoods) and new housing construction in the outer suburbs many people will find it hard to believe the population of the entire urbanized area around St. Louis declined by 2 percent in the 1970's.

In spite of the fact that they are losing population, many older cities and urbanized areas are gaining households. The number of households in the St. Louis urbanized area increased by 12 percent in the 1970's. What makes such apparently contradictory developments possible is the fact that the average number of members of a household is declining. The fertility of married women has decreased, and changes in living arrangements have been brought about by divorce, separation and the virtual disappearance of the extended family living in one dwelling. All these factors have contributed to a decrease in the size of the average household.

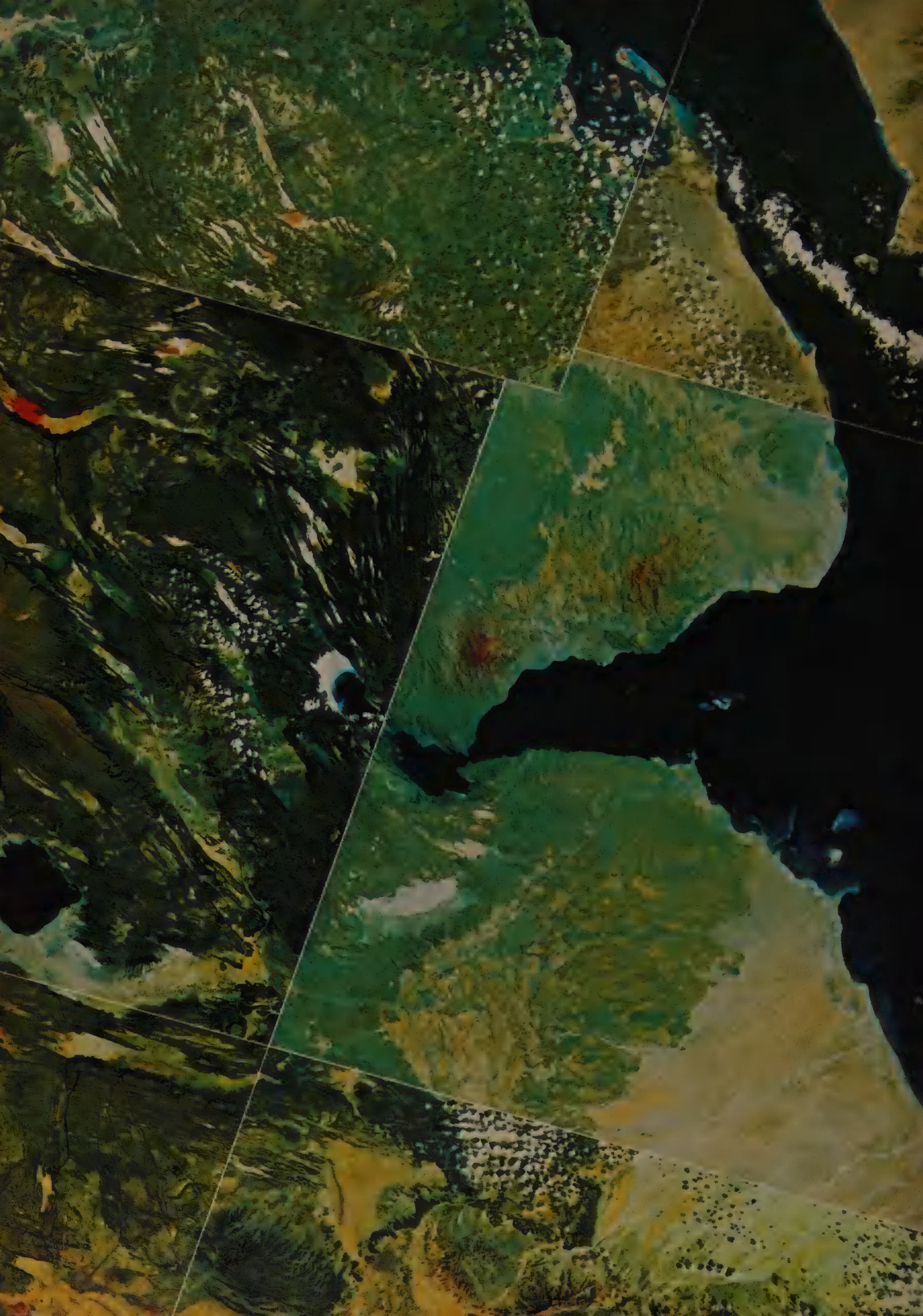
For this reason some of the most significant demographic trends of the past three decades in urban areas have been largely invisible. A count on a large scale, such as the census, is needed to measure them precisely. Our purpose in this article has been to analyze such trends by employing definitions that can be applied consistently over the entire period. The definitions themselves, however, do not remain static. Based on the results of the 1980 census, demographers at the Census Bureau have already begun to redefine the urbanized area. Urban areas are no longer limited to the territory around cities with a population of 50,000 or more. As of 1980 an urbanized area is any group of 50,000 people with an average density of 1,000 per square mile, regardless of the size of the largest city. Thus many people who live in or near a small town have become part of the urbanized-area population.

The change is not particularly surprising. Almost every redefinition of what is urban has expanded the number of people in the category. Indeed, as urbanization reaches its historical peak some uncertainty about who is an urban dweller is almost inevitable. The inability of demographers to agree on what is urban territory may be taken as a sign that urbanization has peaked at a high level.

The uncertainty undoubtedly extends even further than we have proposed here. For some time almost the entire population of the U.S. has been urban in the sense of having access to the amenities of urban life whether or not they live in cities. Perhaps the best model of the population is a continuum in which one end represents the most nearly urban part and the other represents the most nearly rural part. In the 1970's there was an unexpected increase in the demographic growth of the most nearly rural parts of the nation; at the same time the rate of loss increased in the most nearly urban parts. The fundamental cause of the reversal appears to be economic: population has followed jobs. The future distribution of population in the U.S. could well be determined by which regions are able to compete for new jobs both within the U.S. and in international markets.



ST. LOUIS URBANIZED AREA shows a pattern that prevails in many older cities. The boundaries of the city proper (*orange*), the inner suburbs (*purple*) and the outer suburbs (*blue*) are shown on the map at the left. The rate of population growth in each area is shown in the chart above. The city of St. Louis has been losing residents for decades and now has the same population it had in 1890. Many people who left the city moved to the surrounding suburbs. Such a pattern is often described as a doughnut. In the 1970's the area of population loss (the "hole" in the doughnut) spread outward: the inner suburbs lost a substantial fraction of their residents for the first time. The outer suburbs are still gaining population, but at a rapidly decreasing rate.



How Continents Break Up

They rift apart over millions of years, and in the process they deform. A study of the rifting is beginning to reveal the properties of the plates that compose the earth's crust

by Vincent Courtillot and Gregory E. Vink

The theory of plate tectonics holds that the crust of the earth consists of plates some 100 kilometers thick that are moving with respect to one another. It accounts for a wealth of facts. The distribution of earthquakes and volcanoes around the world is largely explained by the motions of the plates: their creation at midocean ridges; their destruction by collision or at subduction zones, where they plunge downward into the earth's mantle; their friction as they slide past each other along what are called transform faults. Moreover, the similarity in shape between the edges of two continents that today are separated by thousands of kilometers (for example the eastern edge of South America and the western edge of Africa) shows that they rifted apart. In other words, they were once on a single plate. Corroborating evidence is supplied by the distribution of plant and animal fossils on the continents.

If it is assumed that plates do not deform as they move, their relative motions can be calculated by means of mathematical theorems describing the motion of a rigid body on a sphere. Indeed, the motions of most of the earth's large plates over the past 150 to 200 million years have been worked out with considerable precision. On the other hand, the motions of plates that are assumed to be rigid cannot yield more than an understanding of their kinematics. The assumption that plates are rigid means it must also be assumed that their boundaries have a fixed geometry, and this latter assumption makes it impossible to propose a realistic model of how continental crust breaks up. For example, the breakup that created South America and Africa must be taken to have happened instantaneously along a line, a most unlikely event.

Actually there is ample evidence that continental crust does deform. Mountain ranges, which often result from the collision of plates, are the most obvious evidence. In addition the distribution of earthquakes at plate boundaries suggests that the boundaries are not narrow, and the occurrence of earthquakes well away from plate boundaries suggests the release of stress, and thus the occurrence of deformation, in the middle of a plate. The work we shall describe in this article suggests that the breakup of continental crust must be seen as a process taking millions of years along a zone several hundreds of kilometers wide. It suggests, moreover, that the crust at such a zone is unlikely to behave like anything rigid.

The key evidence supporting the theory of plate tectonics is found on the ocean floor in the form of magnetic anomalies. Fundamentally the volcanic material that rises into place to become new oceanic crust at the axis of a midocean ridge is hot enough for the minute magnetic domains in it to become aligned with the magnetic field of the earth. As the material cools, locking the alignment into place, it spreads away from the ridge axis. Behind it along the axis newer crust is rising into place. Now, the magnetic field of the earth varies in an irregular pattern, and its polarity reverses approximately five times every million years. Therefore as the new oceanic crust spreads away from the ridge axis two divergent magnetic "tapes" record the pattern of reversals. This makes it possible to measure the age of the ocean floor and also determine the velocity at which plates have diverged. That velocity proves to vary from less than a centimeter per year to more than 15 centimeters. Moreover, when the ridge axis is offset by a series of transform faults, the magnetic pattern in the oceanic crust should reproduce the offset.

There are instances, however, where the magnetic pattern fails to match the offset. These mismatches were initially explained in terms of sudden jumps of the ridge to a new location. The problem is that some of the jumps are so numerous and so closely spaced in time that the process is best seen as the continuous evolution of the boundary of a plate. Consider the peculiar *V*-shaped magnetic patterns found in the Pacific, first near the Galápagos Islands and then on the Juan de Fuca Ridge off the Pacific Coast of North America. A study of those patterns led Richard N. Hey, who was then working at the Hawaii Institute of Geophysics, and his colleagues (particularly Frederick K. Duennebier of the Hawaii Institute, W. Jason Morgan of Princeton University and Peter R. Vogt of the U.S. Naval Research Laboratory) to develop a model in which a rift propagates continuously through the oceanic crust at a rate comparable to the rate at which plates spread apart at the rift.

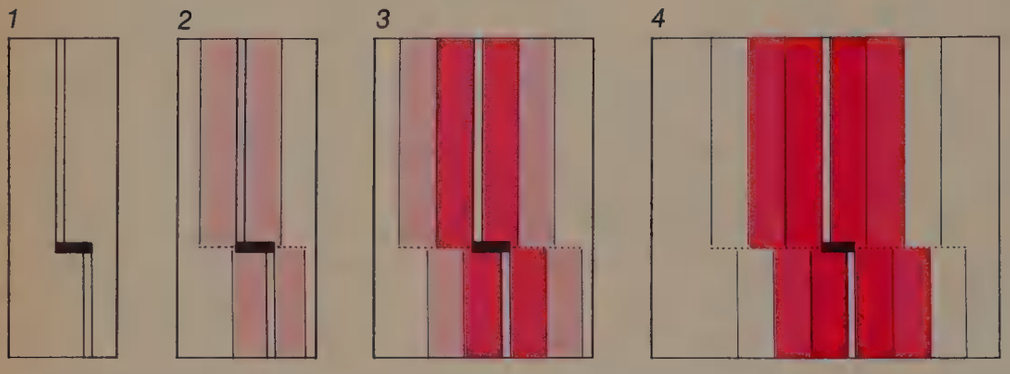
As Hey and Tanya M. Atwater of the University of California at Santa Barbara explain it, the process occurs where two rifts are offset by a transform fault. If the process is modeled in steps, each step is a sudden jump in which the tip of the rift propagates forward, establishing a new fault and moving oceanic crust from one side of the rift (that is, from one plate) to the other [*see illustrations on next page*]. The result is a *V*-shaped

AFAR DEPRESSION in northeastern Africa is one of two places on the earth (the other is Iceland) where a midocean ridge comes up onto dry land. The axis of each such ridge is a rift where two great plates diverge and new oceanic crust rises into place. The positions of the rifts at Afar are controversial. According to the authors, one rift courses westward from the Gulf of Aden into the Gulf of Tadjurah, the beak-shaped body of water extending into the image from the right. Its tip, currently at Lake Asal just west of the Gulf of Tadjurah, is propagating into the depression and will ultimately join with a rift whose tip is moving southward in the Red Sea. (The mouth of the Red Sea is at the upper right.) The rifting will then have given rise to a new ocean basin. The image is a mosaic of six frames made in false colors by the Landsat satellites.

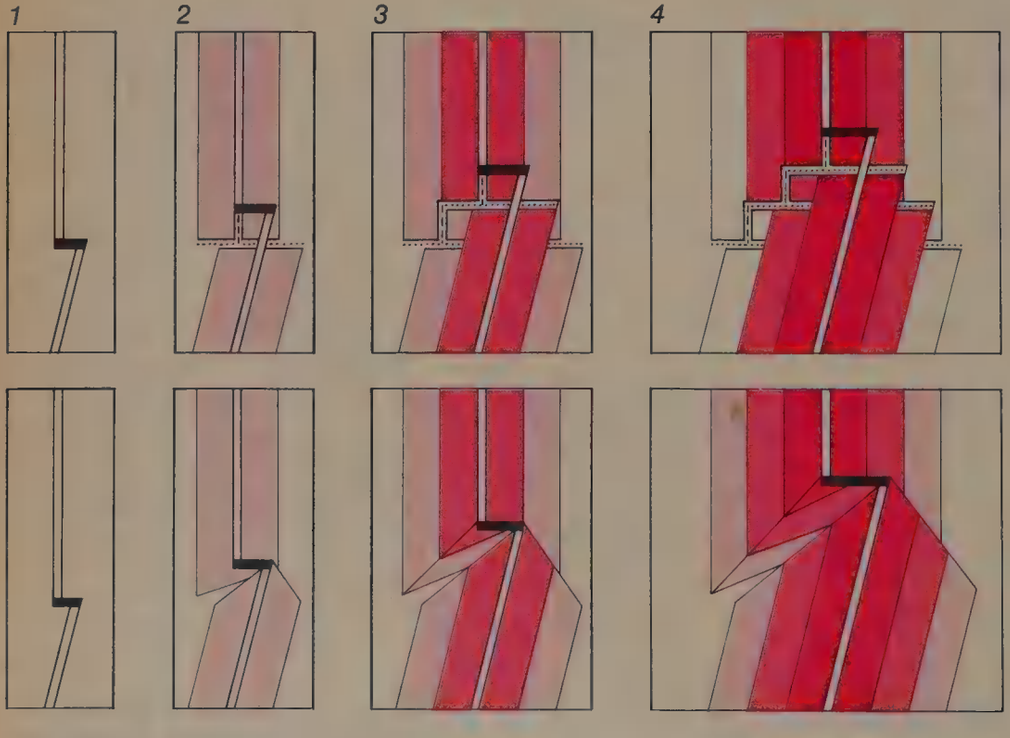
wake consisting in part of small pieces of crust bounded by abandoned faults. If the process is taken to be continuous, as it in fact is in the crust of the earth, the wake is a *V*-shaped discontinuity in the pattern of magnetic anomalies. Hey pointed out that the discontinuity can mistakenly be attributed to a fracture. (The attribution is mistaken because no slippage between plates ever occurred along the discontinuity.) He called each limb of the *V* a pseudofault.

The breakup of continental crust must also involve the propagation of rifts, yet the details of the process have been elusive. In the early 1970's J. Tuzo Wilson of the University of Toronto and Kevin C. Burke and John F. Dewey of the State University of New York at Albany suggested that there might be a link between "hot spots," triple junctions and continental breakup. They suggested that if a continent becomes stationary with respect to the underlying mantle, thermal anomalies (hot spots) in the mantle might make the continent dome upward and rupture in a three-branched pattern of rifts. (Doming and rupture often seem to give rise to a three-branched pattern. They do it, for example, in the crust of a pie baking in the oven.) The three rifts might then develop into two active rifts and an aborted one. (Burke and Dewey thought they could see examples along the rifted margins of the Atlantic.) The active rifts propagating away from a number of junctions could eventually join up and thereby fragment the continent.

More recently two other proposals have been made. First, Morgan has proposed that rifting often occurs along a hot-spot track. Here the motion of a continent over a thermal anomaly in the underlying mantle gives rise to a line of volcanoes at the surface. Under the surface, meanwhile, the thermal anomaly thins and weakens the crust, making it prone to rift. Second, Hey and Vogt have suggested that a process resembling the model of a rift propagating through oceanic crust might account for some of the rifting on the continents. On the ocean floor, however, the model requires that one rift recede while another one advances. Thus the model can only modify the geometry of preexisting plate boundaries. It cannot apply to the breakup of a continent, a process in which a new plate boundary arises.



MAGNETIC ANOMALIES on the ocean floor are the strongest evidence that new oceanic crust spreads away from each midocean rift. The rift itself (*1*) is offset by a transform fault (*black*). New crust arises continuously. At first it is hot; thus its intrinsic magnetization becomes aligned with that of the earth. As it cools and solidifies, the magnetization is locked into place; hence at each side of the rift a band of magnetized crust diverges (*2*). The earth's field occasionally reverses its polarity. Each such reversal entails two new magnetic anomalies (*3, 4*).

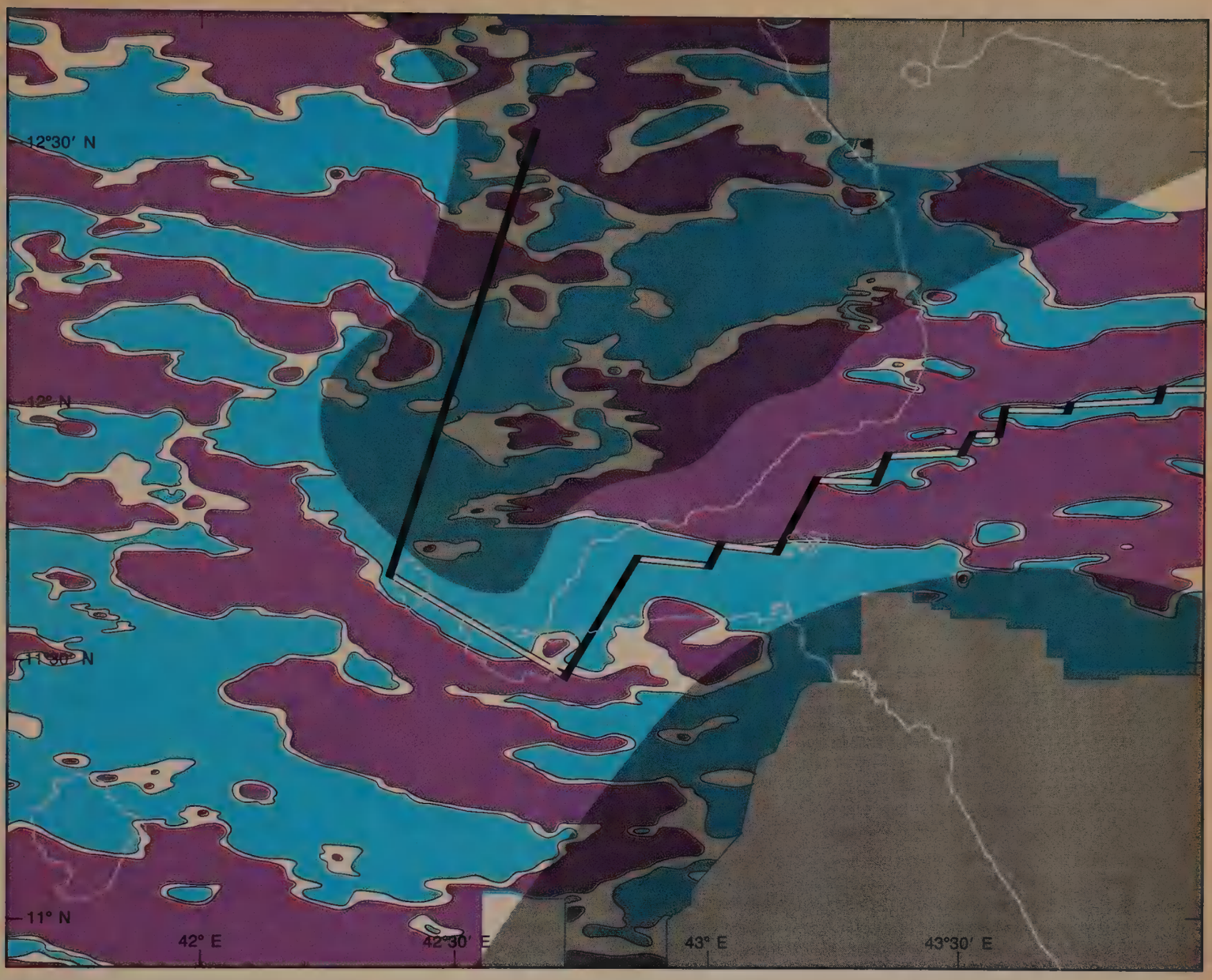


DISRUPTED PATTERN of magnetic anomalies is evidence that rifts propagate. In the top row a model proposed by Richard N. Hey of the University of California at San Diego and his colleagues is shown in steps. Two rifts at an angle to each other are offset by a transform fault (*1*). New oceanic crust rises into place and spreads; then the tip of the southern rift propagates northward, transferring a block of the new oceanic crust from east to west of the rift, that is, from one plate to another (*2*). Successive episodes of sea-floor spreading and rift propagagation (*3, 4*) produce a *V*-shaped zone of magnetic discontinuity. In the bottom row of the illustration the process is continuous; thus the sea floor spreads while the rift is propagating.

At this point magnetic anomalies are again important evidence. In 1977 one of us (Courtillot), working with Jean-Louis Le Mouël and Armand Galdeano of the University of Paris, made a magnetic map of part of the western Gulf of Aden and part of the adjacent land mass, the Afar depression in Djibouti and Ethiopia. Afar is a natural laboratory where sea-floor spreading can be studied without the hindrance of thousands of meters of overlying water. Like Iceland, it is an exposed part of the worldwide midocean-ridge system. Many investigators believe much can be learned about the fundamental process of rifting in these dry-land settings. As early as 1938 the French paleontologist and philosopher Pierre Teilhard de Chardin was convinced that Afar was the place where Alfred Wegener's theory of continental drift should be tested. More recently the investigators who dived in research submarines to the mid-Atlantic ridge at 37 degrees north latitude as part of the French-American Mid-Ocean Underwater Survey (FAMOUS) were struck by the resemblance of the rift at Afar to what they had seen under almost three kilometers of water.

The clearest clues to what has happened at Afar are the magnetic maps [*see top illustration on opposite page*]. The maps reveal that two types of crust are juxtaposed. In one type the magnetic anomalies are intense and form parallel bands. The pattern is typical of the one that results from sea-floor spreading. The bands can be correlated with the periods between reversals of the earth's magnetic field, and so they can be dated. In the other type of crust the magnetic anomalies are wider, rounder and less intense. They describe a magnetic quiet zone, or MQZ.

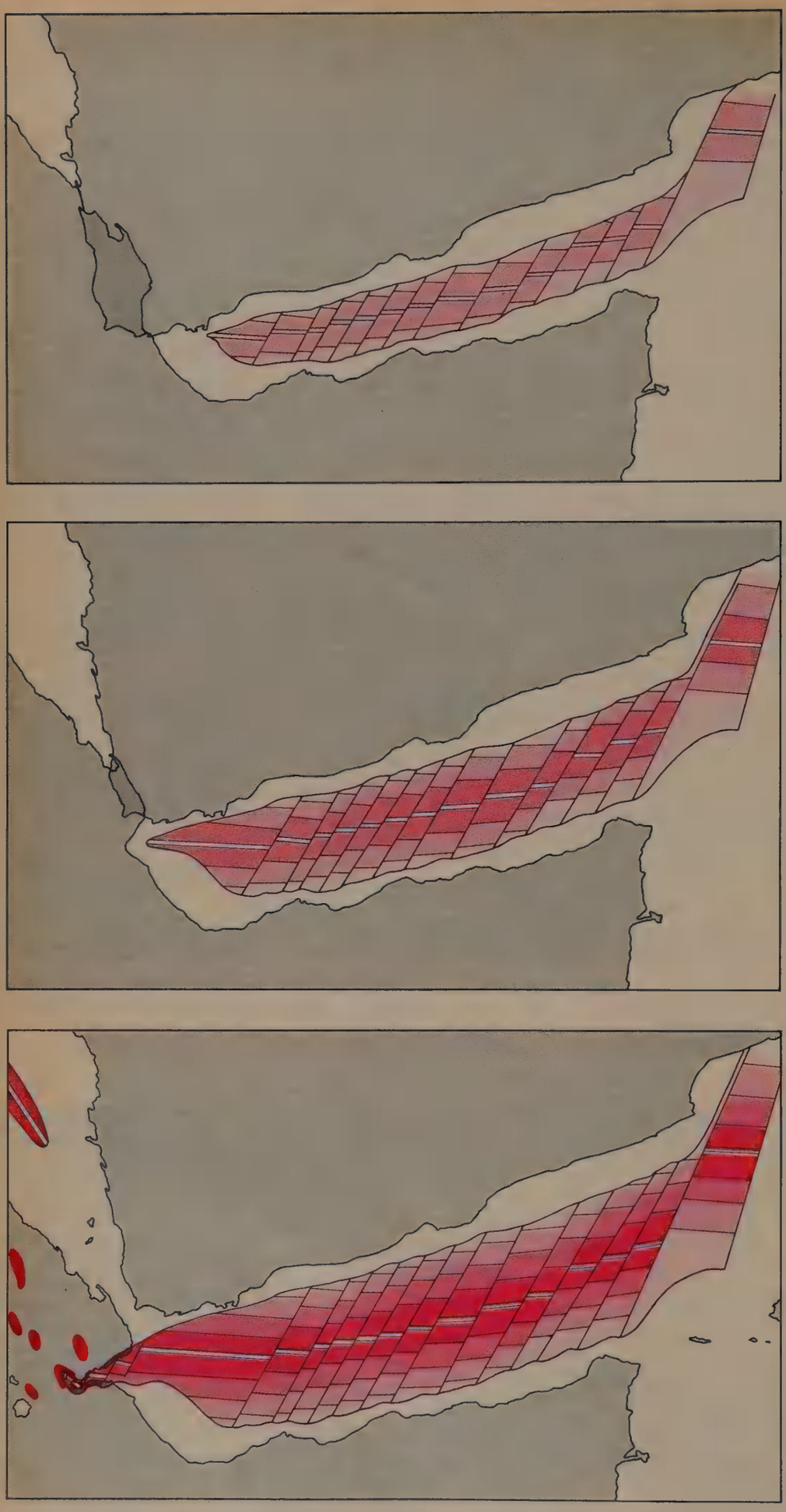
The MQZ forms a huge reclining *V*



MAP OF MAGNETIC ANOMALIES at Afar and on the floor of the Gulf of Aden reveals that different types of crust are juxtaposed. In the gulf the anomalies (shown in alternating blue and red) are intense and elongated, in a pattern typical of oceanic crust. Elsewhere, and particularly toward the northeast, the anomalies are subdued. They mark out a funnel-shaped region called the magnetic quiet zone, or MQZ (*gray*). The MQZ may represent continental crust deformed by the propagation of the rift along the floor of the gulf.



VIEW DOWN THE AXIS of the rift at the Gulf of Aden is afforded to the east from a point on the western shore of Lake Asal, 155 meters below sea level. The rift is at the center. It is flanked on each side by faults running parallel to the rift that produce a set of escarpments. The pattern is much like the one observed at midocean ridges from research submarines under almost three kilometers of water.



RECONSTRUCTIONS OF THE GULF OF ADEN as it looked 10 million years ago (*top*) and seven million years ago (*middle*) are made by taking the present-day map of magnetic anomalies on the floor of the gulf (*bottom*) and removing the anomalies that are younger than those ages. The direction of rift propagation (east to west) is plain. Overlaps between Arabia and Africa (*dark gray*) show that the plates stretched as they rifted. In the present-day map a number of active volcanic ranges (*solid color*) indicate deformation ahead of the tip of the rift. The Red Sea rift is at the upper left of the present-day map. It too is flanked by magnetic anomalies.

that abruptly interrupts the oceanic pattern. It therefore interrupts the bands of magnetic anomalies that mark sea-floor spreading; hence the edges of the MQZ can be dated rather precisely. Between 45 and 46 degrees east longitude in the Gulf of Aden the edges are about 10 million years old; they intersect the magnetic anomaly representing ocean floor of that age. Toward the west the edges become progressively younger. Near Lake Asal in Afar they are very young indeed. A number of earthquakes and the birth of a volcano near Lake Asal in November, 1978, confirm that the vicinity of the lake is a site of tectonic activity today.

The simplest interpretation of these various observations is that a rift is propagating westward into preexisting oceanic and continental crust. The edges of the MQZ were created by past propagation, which formed new oceanic crust in its wake. The Gulf of Aden is the result. The opening angle of the *V* formed by the edges is about 30 degrees. It arises by simple trigonometry from the rate of the sea-floor spreading that has opened the Gulf of Aden (1.5 centimeters per year) and the rate at which the tip of the rift has moved westward (three centimeters per year).

Support for this interpretation has come from a study of seismic activity in the region undertaken by Jean-Claude Ruegg and Jean-Claude Lépine of the University of Paris. They observe that the foci of most earthquakes under the Gulf of Aden cluster in a band approximately 10 kilometers wide that closely follows the rift. (The position of the rift is deduced from a mapping of the bottom of the gulf and from the position of the youngest magnetic anomaly.) The band of earthquake foci extends from the Gulf of Aden into the Gulf of Tadjurah, a beak-shaped westward extension of the Gulf of Aden. After that the band coincides with the Ghoubbet-Asal Rift, which runs from the Gulf of Tadjurah to Lake Asal. Apparently it ends near Lake Asal, where, we suggest, the tip of the propagating rift is currently situated. West of Lake Asal the seismicity becomes diffuse, suggesting that there the crust is deforming over a wider area.

Further evidence comes from the age and the chemical composition of the basalts at Afar, as they have been determined by Michel Treuil of the University of Paris, Jacques Varet of the Bureau de Recherches Géologiques et Minières and Olivier Richard of the University of Paris at Orsay. Basalts are the volcanic rock that composes oceanic crust. The basalts found on the edges of the Gulf of Tadjurah have a tholeiitic composition: they are poor in the mineral olivine but contain the mineral orthopyroxene. Such basalts are formed by the large-scale partial melting of the mantle under a midocean ridge. Closer to Lake Asal

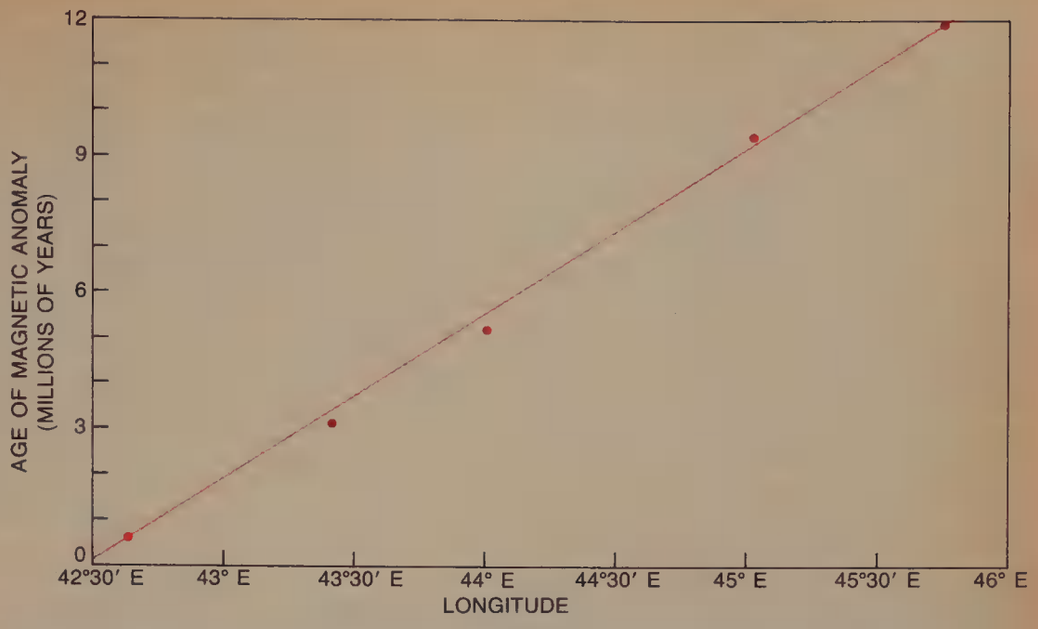
and the tip of the rift the basalts are more alkaline, that is, they are richer in sodium and potassium. This indicates a lesser degree of partial melting than tholeiites undergo. The age of the basalts at each location closely matches the age deduced for the oceanic crust at that location from the measured magnetic anomalies.

In sum, the principal sign of a propagating rift at Afar is a distinctly oceanic pattern of magnetic anomalies that is disrupted by the intersection of the pattern with the rifted edge of a continent. The magnetic quiet zone at Afar is therefore best taken to be continental crust that has deformed in the course of the propagation.

Signs that rifts have propagated elsewhere in the world, so that continents deformed, emerge from reconstructions: the efforts made by earth scientists to determine how continental fragments fitted together before they rifted apart. The usual method is to try to match up their edges. Thus each fragment is rotated about an axis passing through the center of the earth until magnetic anomalies of the same age are made to coincide. If the edge of each fragment arose at a single instant in time, so that the edge is precisely parallel to the oldest magnetic anomalies on the ocean floor that has arisen between the fragments, the reconstruction will succeed: it will match the fragments precisely and at the same time yield the age of the rifting. If, however, the rift has propagated, the reconstruction will not quite succeed: the edge of each fragment will not have arisen all at once, and so the edge will cut across magnetic anomalies.

In such instances bringing together the edges of two continents will not reproduce the configuration they had before they rifted apart. A more accurate reconstruction will match the points where rifting began. The continents will then overlap in places, but that is quite appropriate. The overlaps represent the stretching of the continental crust that took place as the crust was broken apart by propagating rifts.

This appears to be what happened in the Gulf of Aden. If, for example, an attempt is made to reconstruct the relative positions of the African plate and the Arabian plate 10 million years ago (the age of what is called magnetic anomaly 5), it is found that the gulf closes up. Moreover, it is found that the plates begin to overlap at 43 degree 30 minutes east longitude. Ten million years ago the tip of the rift was there. This explains the observation, made by A. S. Laughton and his colleagues at the National Institute of Oceanography in England in 1970 and confirmed recently by James R. Cochran of the Lamont-Doherty Geological Observatory, that anomaly 5 and anomalies older than



SPEED OF PROPAGATION of the rift in the floor of the Gulf of Aden is deduced from the reconstructions. Each point plots the westernmost limit of a magnetic anomaly (*horizontal axis*) and the age of the anomaly (*vertical axis*). The slope of the line connecting the points implies a velocity of three centimeters per year. The end of the line at an age of zero and a longitude of about 42 degrees 30 minutes demonstrates that the tip of the rift is now at Lake Asal.

anomaly 5 cannot be found west of 43 degrees 30 minutes east longitude.

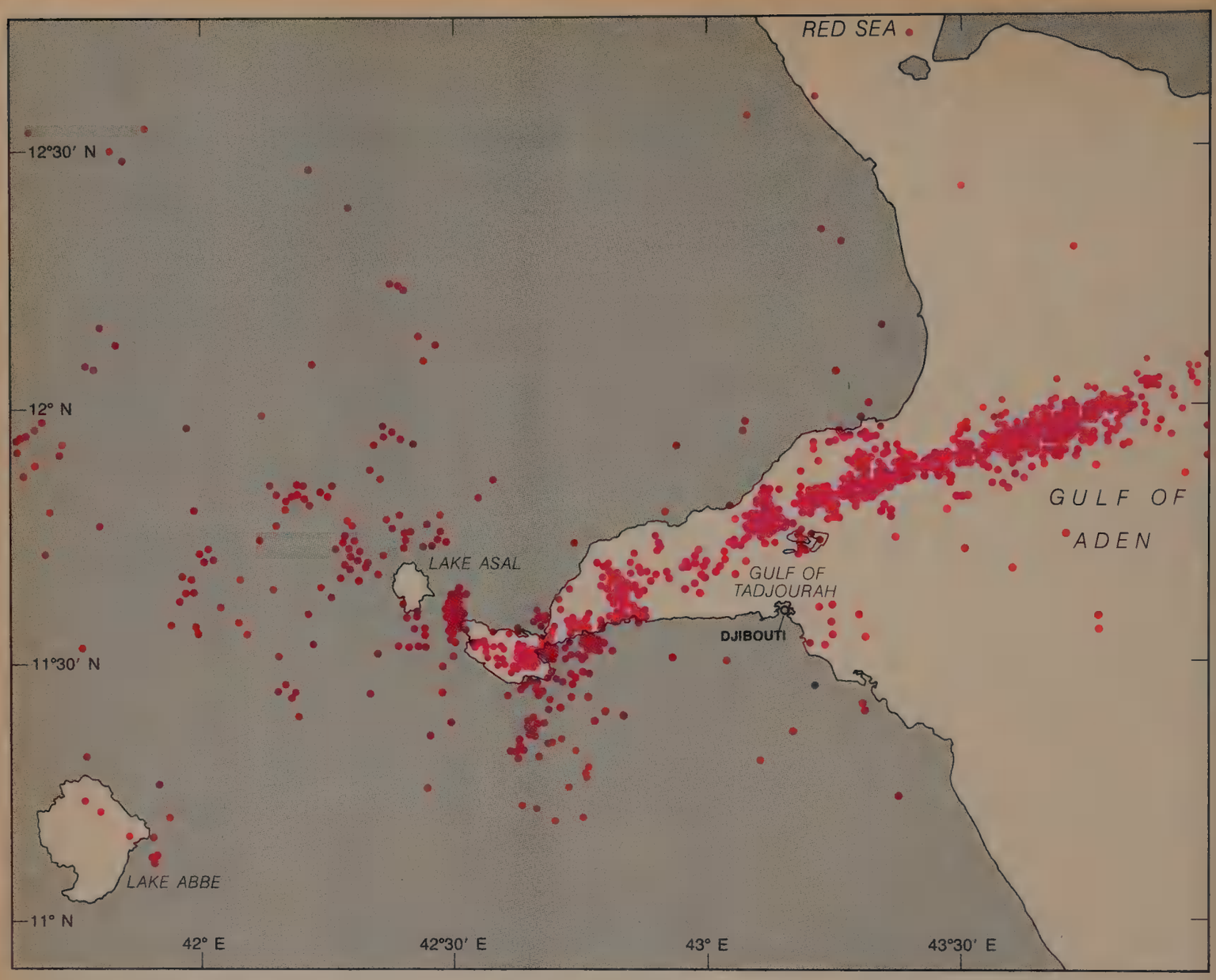
Overlaps also emerge in reconstructions of the North Atlantic made by one of us (Vink). Specifically overlaps between northern Greenland and Spitsbergen suggest that the edge of each of these bodies of land stretched by as much as 100 kilometers. Still another example emerges from the fit of South America and Africa. The classic reconstruction was done by Sir Edward Bullard of the University of Cambridge. Bullard used the 1,000-meter-depth contour lines in the Atlantic to represent the edge of each continent. He rotated the lines together. A computer program found the fit that minimized gaps and overlaps. Even so, there remained a gap of 250 kilometers between the southern tip of Africa and the edge of an undersea plateau that includes the Falkland Islands. The Deep-Sea Drilling Project of the Scripps Institution of Oceanography has since shown that the Falkland plateau is continental crust. It is a submerged part of South America. The gap cannot have existed. When it is closed in a reconstruction, there are overlaps between South America and Africa.

In any case magnetic surveys of the South Atlantic show that the oldest magnetic anomalies on the ocean floor do not parallel the edges of South America and Africa. The surveys confirm that propagating rifts created the South Atlantic. The oldest magnetic anomalies (one with an age of 135 million years near the edge of Africa and one with an age of 127 million years near the edge of South America) are found only in the southernmost part of the Atlantic. The floor of the Atlantic was therefore spreading in the south at a time when Africa and South America were still joined in the north. The rift had not yet propagated that far.

The available evidence suggests the following picture of how a rift propagates to break up a continent and form a new ocean basin. First a system of fractures develops in a plate, possibly along a preexisting weak zone such as an old suture zone (the locus of an earlier collision between continents), a large strike-slip fault system (a fault system where the slippage is horizontal) or, as Morgan has suggested, a hot-spot track. Let us assume for the sake of simplicity that the fractures lie in a linear zone, and let the average width of the zone be 200 kilometers, as it is in the Rift Valley of East Africa. As the fractures develop, the continental crust is stretched and thinned. Nevertheless, the relative motion of the plates that will emerge from the process is negligible: the total stretching is limited to a few tens of kilometers and occurs over periods of a few tens of millions of years.

As the stretching progresses, increasing lengths of the future plate boundary have decreasing strength. The mechanics of the process are unlikely, however, to be uniform. Thus the plates remain attached at a number of "locked zones." These zones cannot prevent the failure of the crust that lies between them. Hence basaltic oceanic crust begins to rise into place. In short, rifting begins. Each rift elongates rapidly through the thinned crust. Then the tips of the rifts enter the remaining locked zones and begin to break them apart.

Let us now consider a single locked zone. As in Hey's model of sea-floor spreading, a sequence of discrete steps is easier to explain than the actual contin-



FOCI OF EARTHQUAKES at Afar were recorded between 1974 and 1980 by investigators from the Institut de Physique du Globe in Paris. They confirm the position of the rift. Scattered foci west of the tip of the rift suggest that the crust there is now deforming.

uous process. In the first step the tip of a rift has reached a locked zone from each side [*see illustration on opposite page*]. Along the length of each rift new oceanic crust is being created. It fills the space freed as the two plates move apart. Inside the locked zone, however, no new crust develops. Therefore the locked zone must stretch.

In the next step the tips of the rifts penetrate an incremental distance into the locked zone. In this increment the crust of the locked zone stops stretching; the creation of oceanic crust takes over. Hence an increment of the locked zone is carried away by each plate. The part of the locked zone not yet invaded by the rifts continues to stretch. The process goes on until the rifts meet and the two plates separate completely. What remains of the locked zone is a deformation in the edge of each newly formed continent. The deformation is maximal at the point where the continents last touched.

The model can be refined so that the process is continuous. It then emerges, naturally enough, that if parameters such as the rate at which the rifts propagate change with time, the deformation of the edge of the continent produces various shapes. For instance, the shape of certain overlaps that arise in attempts to match the edges of continents that broke apart suggests that the propagating rifts may slow down as they first cut through a locked zone and then may accelerate in the final phase of the rupture. (The result in such a case is a deformed zone with a flattened tip.) More generally the model predicts that the strain in a continental plate that breaks up occurs in a band some 100 kilometers wide at the margin of each newly formed plate and also in the locked zones, which stretch dramatically. The disruption of the crust in a locked zone may be responsible for the lack of intense magnetic anomalies in the magnetic quiet zone at Afar.

Yet if Afar is to be viewed as a locked zone in the act of deforming, a second propagating rift in addition to the one that opened the Gulf of Aden must be identified. The East African Rift Valley, which has a long history of very small motion, may be an aborting rift and is not a good candidate. The younger and more active southern Red Sea is more promising. Ron W. Girdler and Peter Styles, who were then working at the University of Newcastle upon Tyne, have attempted to identify magnetic anomalies in the Red Sea so that the seafloor spreading there can be related to the spreading in the Gulf of Aden. The effort has been complicated by the intricate faulted structure of the floor of the Red Sea and by the slow rate at which it is opening. So far Girdler and Styles have been able to recognize two probable phases of spreading, one between 30 and 15 million years ago and then a recent one beginning five million years ago that is continuing today. Magnetic mapping by Hans A. Roeser of the Federal Institute for Geosciences and Natu-

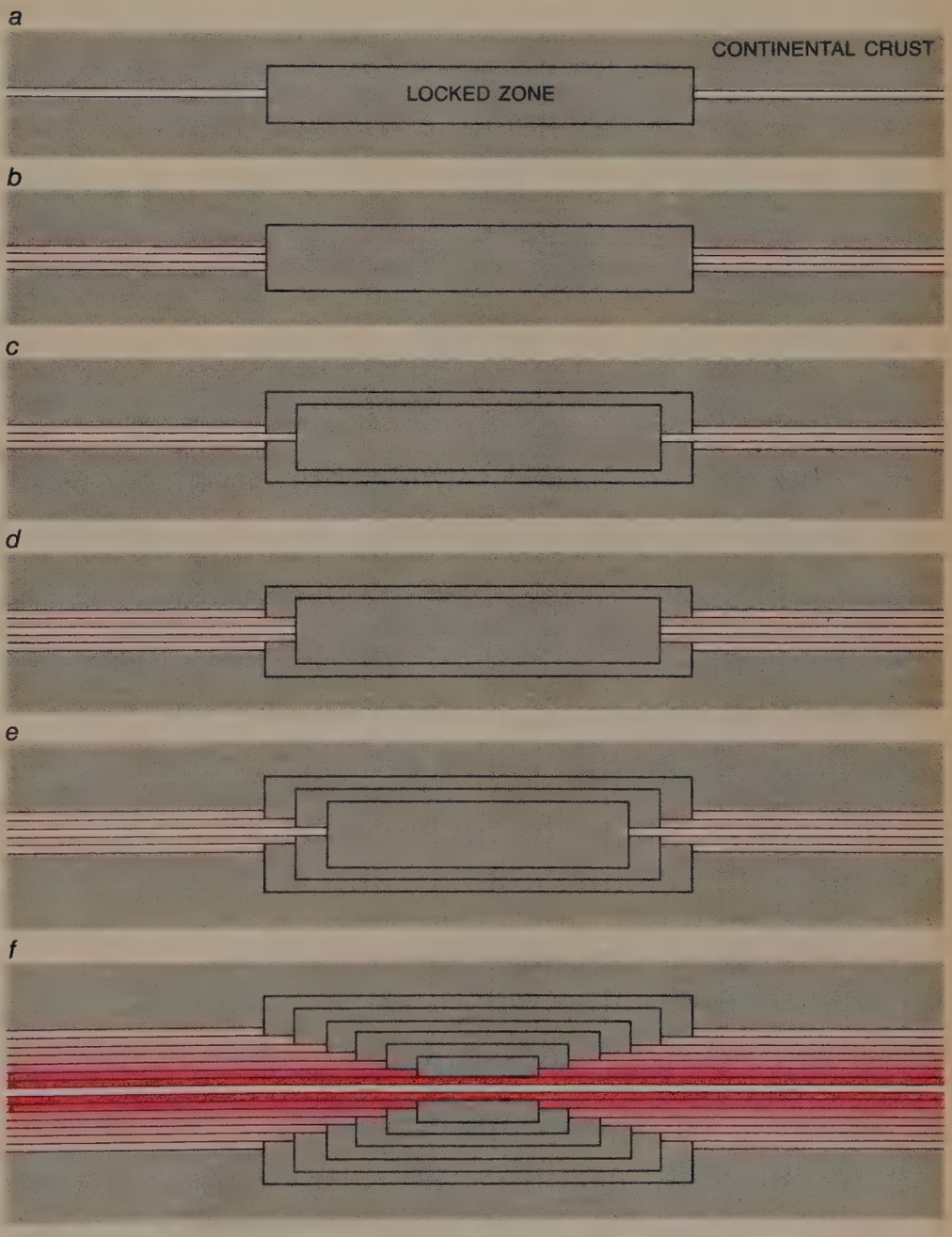
ral Resources in Hanover reveals at the Red Sea a *V*-shaped zone of oceanic anomalies quite like the one in the Gulf of Aden. The magnetic quiet zone at the Red Sea appears to correspond to the older phase of spreading.

The tip of the putative rift that is opening the Red Sea is now between 15 and 16 degrees north latitude. Its distance from the tip of the Asal-Aden rift is 300 kilometers. Presumably that is the length of the Afar locked zone remaining to be broken apart. A comparison of Afar with the model described above suggests that the locked zone was originally 750 kilometers long and 150 kilometers wide. Each of the propagating rifts has therefore torn through 200 kilometers of the zone. An extrapolation of the current rates of propagation suggests that the locked zone will break up completely in less than 10 million years. The zone will by then have undergone a mean deformation that will have stretched it by 80 percent of its original width. The last part of the zone to rupture will have stretched twice that amount.

How do the findings at Afar bear on the suggestion by Wilson, Burke and Dewey that rifting is related to the doming and rupturing of the continental crust above a hot spot in the mantle? Afar has repeatedly been considered one of the clearest cases of a triple junction caused by a hot spot. The junction is that of the Red Sea rift, the Gulf of Aden ridge and the Ethiopian rift (the northernmost part of the East African rift). Our evidence, however, indicates that the rifts at Afar are propagating toward, not away from, an eventual junction. One of us (Courtillot) has suggested that a first phase of tectonic activity at Afar may have generated a knee-shaped rift that opened the Red Sea and left the East African Rift Valley. Preexisting zones of weakness and heterogeneity in the composition of the crust might be responsible for the "knee." A second phase of activity, still in progress today, may then have opened the Gulf of Aden and produced the Red Sea's second episode of spreading. Again a knee-shaped rift appeared, but since the Afar locked zone is not yet cut through, the rifting is not complete. The Afar junction would thus be the superposition of two knee-shaped rifts.

Rifts that are propagating through a continent and are breaking it apart will doubtless be found in places other than Afar. The Gulf of California is a likely candidate, even though its relatively small scale and its great disruption by transform faults make the magnetic data difficult to interpret. The process of rift propagation is well worth understanding. For one thing, the rifting of continental margins and the subsequent deposition of sediments above them govern the evolution of oil-bearing formations. In addition the study of rifting may yield insight into the mechanical properties of continental crust. The pattern of rifting that separated South America and Africa, for example, suggests the locked zones typically have a length of 400 kilometers, a width of 150 kilometers and are spaced at a mean interval of 700 kilometers. The pattern further suggests that sea-floor spreading begins (that is, oceanic crust begins to rise into place at a rift) when about 65 percent of the length of the future plate boundary has lost its strength to resist the breakup.

In attempting to understand how a plate breaks up over a length of thousands of kilometers, we have dealt with temporal scales of no more than a million years and spatial scales of hundreds of kilometers. On such scales we find the continental crust deforming. Among the tools that facilitate the study of how the deformation proceeds two are crucial: the detailed examination of magnetic anomalies and the detailed examination of the edges of the continents. To the extent that the work goes beyond the assumption that plates are rigid it should provide a better view of how a plate evolves.



RIFTING OF A CONTINENT into two separate plates is shown in steps according to a model developed by one of the authors (Courtillot). The process begins along a preexisting zone of weakness. The crust, however, is heterogeneous and so a number of areas resist. One such "locked zone" is shown; it prevents two rifts from meeting (*a*). Along the length of each rift new oceanic crust rises into place; the new crust fills the space freed as the incipient plates move apart. The locked zone, in contrast, must stretch (*b*). The tips of the rifts then begin to invade the zone (*c*). Still more oceanic crust is created. Meanwhile what remains of the locked zone stretches again (*d*). The tips of the rifts move closer together (*e*). Ultimately the rifts meet and the plates separate (*f*). The locked zone has become a deformation in the edge of each new continent. The model can be made into a continuous one in which the size of the locked zone and the velocity of tips of the rifts change with time. Deformations with various shapes result.



The Salmonid Fishes as a Natural Livestock

The remarkable genetic adaptability of this family, which includes salmon and trout, is making it possible to replenish depleted stocks, to establish new ones and to adapt the fishes for "ranching" operations

by Lauren R. Donaldson and Timothy Joyner

Thousands of years ago the salmonid fishes, a group including salmon and trout, flourished throughout the Northern Hemisphere in cool lakes, streams and rivers and in the open sea. In historic times the hand of man has changed this picture through habitat-destroying activities and heavy fishing, with the result that the salmonids have disappeared from many bodies of water and in others their numbers have greatly diminished. Several recent developments, however, offer excellent opportunities for stemming or even reversing the downward trend. The developments rest on the remarkable genetic adaptability of the salmonids and the ability of fish breeders to turn that adaptability to advantage.

The 60-odd species of the salmonid family include thousands of races. Each race is adapted to a unique environment. This diversity can provide fish breeders with an enormous pool of genetic material from which desirable traits can be selected. They include size and shape; rate of growth; the color, taste and texture of the flesh; tolerance of fresh or salt water, and the timing and patterns of migration.

Drawing on these properties, breeders so far have achieved several significant results. Among them are: (1) Salmonids can be induced to "home," or return to spawn after a period at sea, in any cool body of water where they were released as fingerlings, which means that they can be established or reestablished in bodies of water that do not now have them; (2) some salmonid species will adapt to year-round life in fresh water even though they would normally spend part of their lives at sea, and (3) some species can be successfully "farmed" or "ranched" by techniques of aquaculture.

Ranched salmon are homing in large numbers to release sites in North America, Japan and Siberia. Salmon have been successfully transplanted to Chile. The fjords and sounds of the Chilean archipelago have abundant stocks of plankton on which the newly transplanted salmon can feed. In the future carefully selected stocks of salmon planted at the southern tip of South America may be able to migrate to the edge of the Antarctic to feed on the summertime concentrations of krill and deliver that resource back to the release sites splendidly packaged as bright red salmon flesh. Salmon introduced into the Great Lakes have adjusted to that freshwater habitat in a way many people thought would be impossible. Along the rocky coasts of Scotland and Norway fish farmers are raising salmon and rainbow trout to large size in blocked-off sea lochs, fjords and floating pens.

The natural distribution of the salmonids before human beings began to affect it has been fairly well established. In the genus *Salmo* Atlantic salmon (*S. salar*) and European trout species were found around the margins of the North Atlantic and in adjacent parts of the Arctic Ocean. The rainbow-steelhead and cutthroat species of trout were found around the margins of the North Pacific. The chars, members of the genus *Salvelinus,* lived around the margins of both the North Atlantic and the North Pacific basins and in the waters draining into the Arctic Ocean. The six species of Pacific salmon (genus *Oncorhynchus*) were originally limited to the margins of the North Pacific from Taiwan to California, including the Bering and Okhotsk seas and the adjacent parts of the Arctic Ocean.

All salmonids hatch in fresh water and spend at least their first weeks there. Some members of all three genera are anadromous: they migrate to the sea, often for great distances, before returning to their native streams to spawn. Atlantic salmon return to spawn year after year; Pacific salmon return once and die after spawning. Some species of salmonids remain in fresh water all their lives. Several species of Pacific salmon can adjust readily to living permanently in fresh water. Relict populations of cherry and sockeye salmon survive in lakes that have been landlocked for thousands of years.

The homing ability of migratory salmonids is legendary. They swim upstream for many miles, leaping in the air to climb waterfalls and fish ladders, and unerringly make every turn needed to get back to their native stream. Some years ago Arthur D. Hasler and James A. Larsen of the University of Wisconsin at Madison established that the fish home by smell, distinguishing the unique chemical signature of their native stream in concentrations as dilute as parts per billion. The signature is derived from the rocks, the soil and the or-

HOMING SALMON clogging the Tazimina River in Alaska appear in the photograph on the opposite page. They are sockeye, or red, salmon (*Oncorhynchus nerka*), returning to the place where they hatched. They left the river at from two to four years of age, lived for from one to three years at sea and are returning to spawn, finding their way back by the unique chemical "smell" of their native stream. The fish in ranks at the edges of the stream are resting before continuing upstream. The rank formation arises because they orient themselves upstream; they are side by side owing to their numbers. The other fish are making nests in the gravel. The photograph was made by Thomas C. Kline, Jr., of the University of Washington as part of an annual survey designed to forecast for commercial fishermen and conservation authorities how many sockeyes can be expected when the fish hatched from the eggs eventually return.


CHINOOK
SOCKEYE
COHO
ATLANTIC SALMON
STEELHEAD TROUT

ganic matter associated with the water.

Attempts at artificially rearing the salmonids started in Germany before anything was known about homing except that the fish did it with great success. In 1763 it was observed that the eggs could be squeezed out of the females, fertilized with sperm from the males and incubated in cool, flowing water until they hatched. The early investigators worked with the brown trout (*Salmo trutta*) and the Atlantic salmon (*S. salar*). In the U.S. the effort to culture the salmonids began in 1804 with the artificial spawning and incubation of the eastern char (*Salvelinus fontinalis*).

The first hatchery for Pacific salmon was built on the McCloud River in California in 1870. As in all previous efforts to propagate these fish, the river was robbed of its wild stock by seining the spawning fish from their ancestral spawning grounds to provide the "seed" for the hatchery. The same practice was followed in all the hatcheries established later along the entire Pacific coast from California to Japan. The hatcheries concentrated solely on the incubation of eggs, giving no thought to the problems of feeding and rearing the fry, which were released unready for the challenges of a hostile environment. As a result the early efforts contributed little to the salmon runs and in fact often impaired them.

In 1895 the state of Washington built a hatchery on the Kalama River, a tributary of the Columbia River. This station concentrated on the chinook (*Oncorhynchus tshawytscha*), the largest of the Pacific salmon. At the station procedures were developed for spawning, hatching, rearing, releasing and recapturing the fish to be spawned to provide eggs for the next generation: a spectrum of actions that made it possible to recharge the salmon runs. As the system developed it began to yield surplus spawn that could be transferred to other stations to start new runs.

In 1901 a salmon hatchery was built on the Green River, a small river that enters Puget Sound in the Seattle harbor. Chinook eggs from the Kalama hatchery were brought in to start a new run. After a period of adjustment these chinooks began homing on the Green River hatchery in such numbers that a surplus was soon available. The Green River hatchery became a source of chinook eggs for many other hatcheries in the Puget Sound area, including (in 1949) the experimental hatchery on the campus of the University of Washington. These chinooks, along with the stocks of rainbow and cutthroat trout and coho salmon already established there, provided the university with its own "herds" of salmonids and also provided the basis for a number of experimental projects.

Experimenters elsewhere had for some time been concocting specific recipes of chemicals to be put in the water where the young salmonids were released in order to ensure that they would return there. At the University of Washington we soon learned that such a step is unnecessary; the natural chemical signature of the water is sufficient. A major project entailed selective breeding to increase the proportion of chinook salmon that return to the university pond after two or three years at sea in relation to the number returning after four or five years. This project has been highly successful. So has a project for accelerating the growth of juvenile coho salmon by giving them special diets and controlling the water temperature. Under this regime our coho are able to smolt and go to sea after only six or seven months, compared with the normal period of from 18 to 30 months. These stocks and the techniques employed to rear them are now becoming the basis for commercial salmon ranching and farming in the U.S., Canada, Chile and France.

The chinook salmon brought to the university had to adjust to an environment far different from the environments commonly frequented by this species. Instead of coming back to big, deep rivers such as the Sacramento, the Columbia, the Fraser and the Yukon, to which many of their relatives are accustomed, chinooks returning from the sea to their "home" at the university must enter Puget Sound, turn left, enter the Lake Washington ship canal and pass through the locks either with the ship traffic or by way of the fish ladder along the south bank. Then, after a three-and-a-half-mile trip through the congested industrial area along the shores of Lake Union, they must turn left again, climb a small ladder and enter a collecting pond on the campus.

The attraction water that guides the fish into this last step of their journey is pumped from the ship canal into the collecting pond, from which it flows by gravity back into the canal. This tiny flow has served as the home stream for many generations of salmon and sea-run trout bred at the university. The university's chinook salmon have adjusted to this unusual environment so well that each fall the return to the campus, even after the chinooks have run the gauntlet of heavy sport and commercial fishing in the ocean, the Juan de Fuca Strait and Puget Sound, yields from 10 to 20 times the number of eggs needed to perpetuate the run.

The surplus eggs have served to build many other salmon runs. From 1967 to 1970 several hundred thousand eyed eggs (eggs in which the developing eyes can be seen, the earliest stage at which eggs can be handled carefully without damage) were sent to Michigan as seed for stocks that have become the basis for the lively new sport fishery for chinooks that has developed in Lake Michigan. In 1970 eggs were shipped to the New York State Department of Environmental Conservation for seeding Lake Erie and Lake Ontario with chinooks.

In 1980 eyed eggs from the university's stock of chinooks were sent to Chile. The young fish reared from those eggs were released into waters leading to the Southern Channels zone of that elongated country; there an abundance of crablike and shrimplike plankton is an ideal food source for foraging salmon. Two years later the Chilean and U.S. sponsors of the project were delighted by a return of adult chinooks that far exceeded their modest expectations. These fish are now providing a source of eggs for an expanded planting program that will soon include releases into the Strait of Magellan. Salmon venturing out to sea from those waters should have access to the enormous stocks of krill that congregate in the summer in the region of the Antarctic Convergence, where waters flowing outward from the Antarctic meet warmer waters from the ocean basins to the north. The fish should also be able to forage in the plankton-rich waters of the Patagonian Shelf, which might be more hospitable for them during the winter.

Chinooks are not the only salmonids whose numbers have been increased by these techniques. Programs also have been applied with success to the coho salmon (*Oncorhynchus kisutch*), the pink salmon (*O. gorbuscha*), the chum salmon (*O. keta*), the Atlantic salmon (*Salmo salar*) and the rainbow trout (*S. gairdnerii*).

Coho salmon were historically distributed from the Arctic to central California. Since they tend to spawn and rear their young in small streams, their homing journeys scattered them, and unique stocks or races of cohos evolved in each of thousands of spawning areas

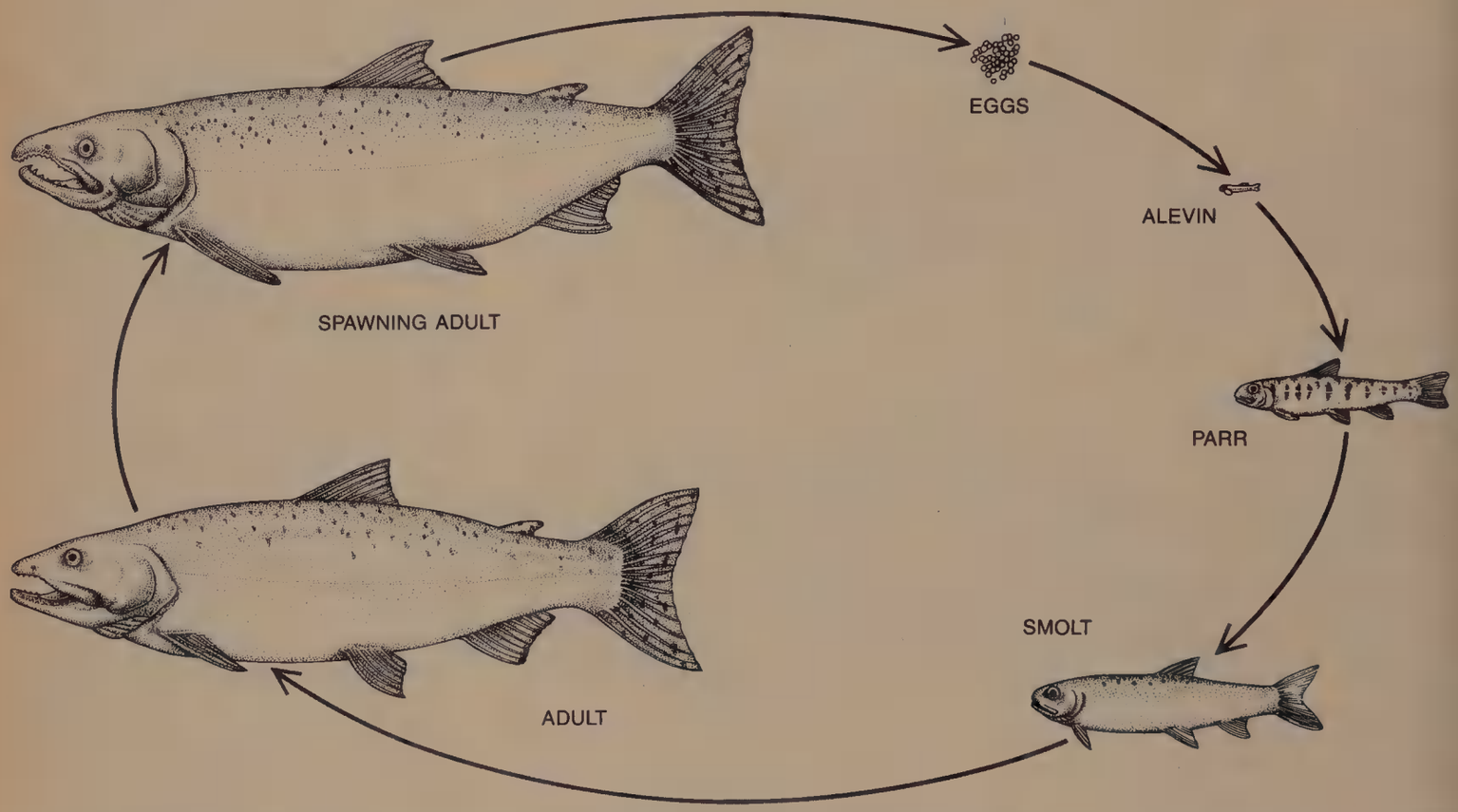
FIVE SPECIES of salmonid fishes, a family that includes the salmon and the trout, are portrayed on the opposite page at a common scale. All are males; at the right they are shown at spawning age and at the left as young parr, still in the freshwater stream where they hatched. Not long after the parr stage the fishes become smolts, that is, they go through the physiological changes that make it possible for them to migrate to sea and live in salt water. The top three species are Pacific salmon, which return to their native stream to spawn only once and then die. The Atlantic salmon and the steelhead trout return to spawn every year for many years.

in streams and tributaries with access to the North Pacific.

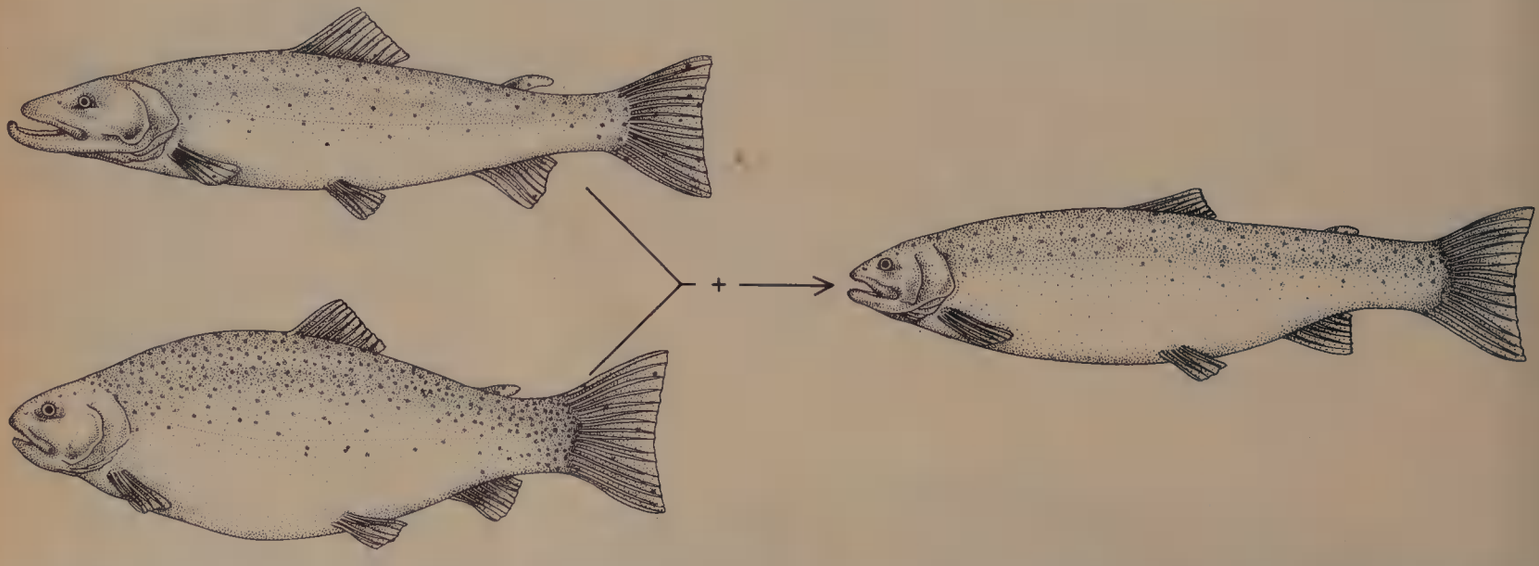
The adult cohos return to their home streams from their feeding areas in the ocean late in the fall, when the seasonal rains increase the flow and reduce the temperature of the water. In the long winter months the eggs lie covered with protective layers of gravel. They hatch between March and June. During the summer the young cohos feed on aquatic insects in the small streams where they were hatched. In the fall, when the adults return to spawn, the young from the previous spring feed on the food organisms dislodged by the spawning fish as they dig nests. The young also eat any eggs that are not immediately covered with gravel.

During the second winter the young cohos remain in the stream, virtually dormant. In their second spring the young in the southern part of the coho range become smolts: they go through the physiological changes that enable them to migrate to sea. (Without these complex changes the salmon, which have a salt content of about 12 parts per 1,000 in their blood, would dehydrate on entering seawater, which has a salt content of from 30 to 32 parts per 1,000.) Cohos stay at sea for a year and a half, returning at the age of three to start the cycle anew. In the colder parts of the range cohos usually need two and a half years to reach the smolt stage; after a year and a half in the ocean they return as four-year-old adults.

Coho salmon have been the "bread



SIX STAGES in the life of the coho salmon (*Oncorhynchus kisutch*) are depicted at a common scale. They are: the eggs; the alevin, to which the yolk sac is still attached; the parr, a spotted juvenile still living in fresh water; the smolt, a silvery fingerling ready to migrate to sea; the mature adult ready to return from the sea, and the mature adult coho after it has returned to its native stream to spawn.



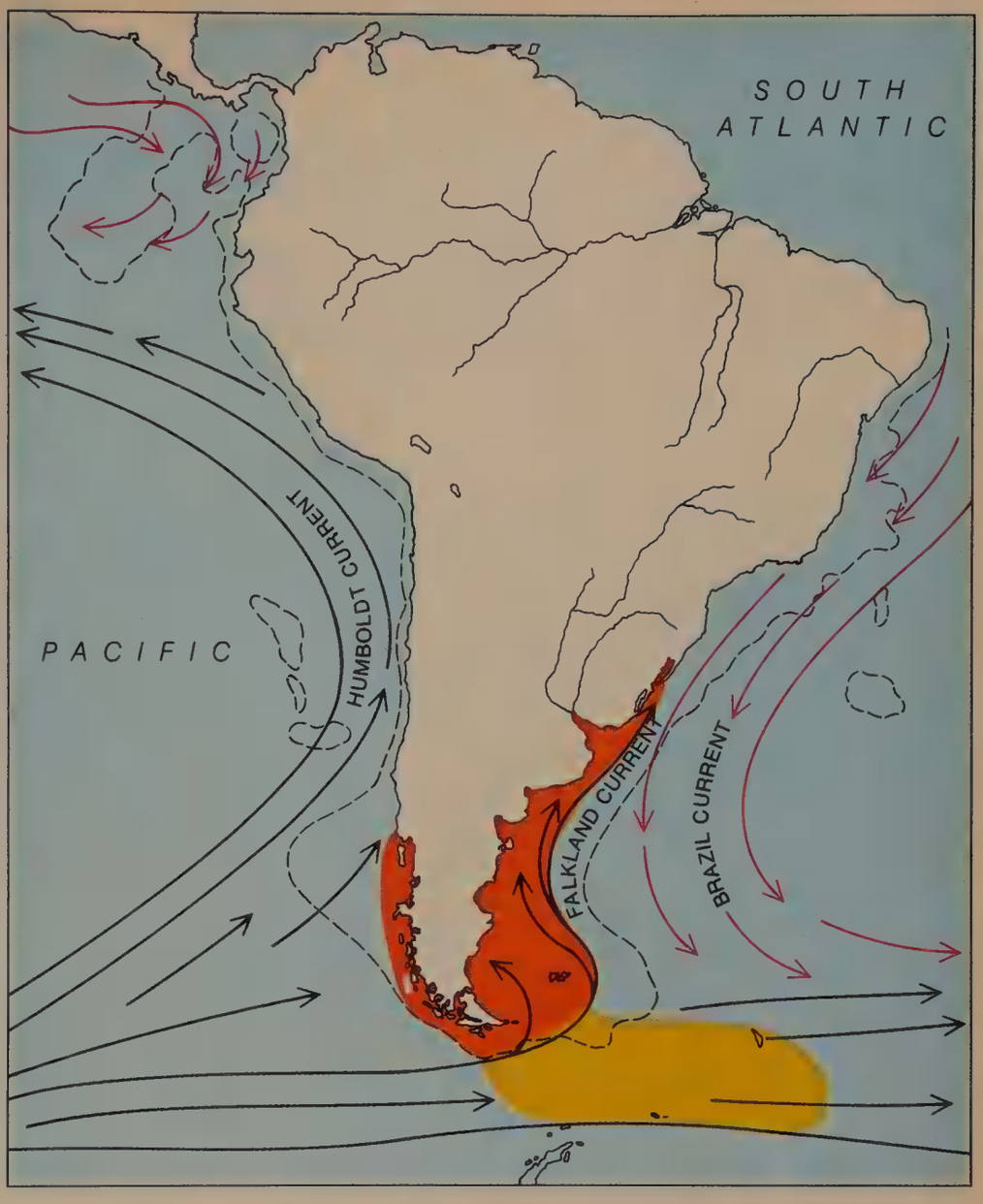
HYBRIDIZATION OF TROUT was carried out at the University of Washington. Sperm from steelheads, which are anadromous (migratory) rainbow trout, served to fertilize eggs from a select stock of nonmigratory rainbows that had been developed at the university. Here the male steelhead is at the top left. The resulting hybrid grows faster than the steelheads but slower than the nonmigratory rainbows. It also migrates, and on its return it displays the rainbow's trait of biting at a lure, which makes it a good fish for sport fishermen.

and butter" fish for salmon hatcheries on the West Coast of the U.S. The millions of coho smolts turned out each year by Federal, state, Indian and private hatcheries now constitute the largest single source of salmon for the commercial and sport fisheries in California, Oregon and Washington. In 1967 one of us (Donaldson) organized a small research project at the University of Washington aimed at determining whether the growth rate of coho salmon could be accelerated with improved diets and warmed water so that fully smolted fingerlings would be ready to migrate in six or seven months instead of 18. We found that the objective was indeed attainable. Most of the cohos released as accelerated smolts returned to the university pond after 18 months at sea, that is, when they were two years old. By selecting brood stock from this group we were able to increase the rate of returning adults from 1.25 percent in 1969 to 6.42 percent six years later.

That improvement resulted from our application of recent research findings on the nutrition of the salmonids by developing diets with appropriate amounts of digestible protein, fats, minerals and vitamins. Fish meal has been the major source of the protein, but the quality of the product from Peru has been poor and supplies from the Northern Hemisphere are becoming both expensive and scarce. Yeast, vegetable protein and bacterial protein have been substituted in various amounts with various degrees of success. Fish-processing wastes preserved by acidification hold much promise.

Heat as a by-product of the generation of electricity or other industrial processes can be employed to hold the temperature of hatchery water in the optimal range for the growth of salmonids (usually from 8 to 12 degrees Celsius, or from 46 to 54 degrees Fahrenheit). Several advantages are gained by controlling the water temperature to accelerate the rearing of hatchery smolts: harmful extremes of temperature in the summer and winter can be avoided; smolts can be released before water levels get dangerously low in the summer; exposure to disease is reduced by the shorter residence time of the fish in the crowded conditions of hatchery ponds; less food is needed with shorter rearing times, and the time to harvest is reduced.

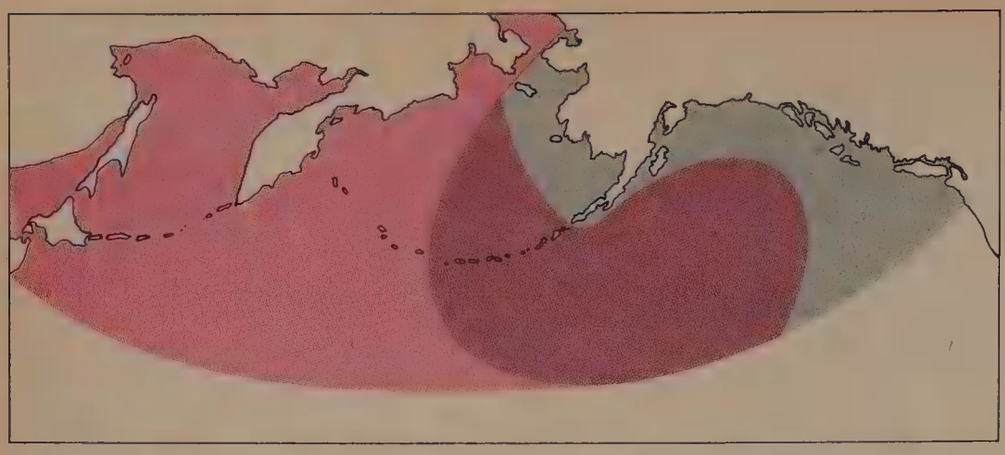
Many of the natural spawning and nursery areas for salmon have been destroyed in the developed nations of the Northern Hemisphere by the cutting down of forests, by farming and by the spread of urban-industrial complexes. Several countries have turned to artificial production to replace the salmon lost by the destruction of habitats. Over the past 20 years the Japanese have had remarkable success in producing chum salmon at hatcheries. Nearly all the chums now harvested in Japanese territorial waters are of hatchery origin.



NEW OCEAN RANGE has been opened to Pacific salmon as a result of the transplantation of fertilized eggs to Chile from North America and Japan. The most successful transplant was made in 1980 from the University of Washington's stock of chinooks. Returns from the stock are providing a source of eggs for an expanded planting program that will release salmon into the Strait of Magellan. The map shows prospective ranges for salmon in the sub-Antarctic.

Although a much greater variety of methods for artificially rearing many species of salmon and trout in hatcheries have been developed in North America and Europe, none of the methods has been applied with the single-minded intensity that has characterized the Japanese application of a system imported from New England in the 18th century, nor have the results been as spectacular. Nevertheless, out of all these efforts a new concept of ocean ranching of salmonids has begun to emerge. It differs from earlier hatchery techniques in that the young fish are carefully nurtured until their prospects for survival in the ocean are optimized. Only then are they released to migrate to the sea, where they can feed and mature on the open range until they return to their native coasts and rivers to be harvested. Ocean ranching will have a profound effect on the evolution and survival of salmon and sea-run trout. If it is managed well, the effect will be beneficial; if it is managed badly (so that, say, inappropriately selected stocks are put into a coastal ecosystem where they usurp habitats of other valuable stocks), the effect could be detrimental to both the ecosystem and the rational management of the salmonid fisheries.

In Oregon, where the ocean ranching of coho, chinook and chum salmon is



LONG MIGRATION of chum salmon (*O. keta*) reared artificially in Japan and the U.S.S.R. takes them to the Gulf of Alaska. The Asian chums are represented by the colored area, the North American chums by the gray area, and the region where they mingle by a mixture of color and gray. On their annual return to Asia the chums yield a harvest of 200 million pounds.

legal under close supervision by the state, a subsidiary of the Weyerhaeuser forest-products company has built a large hatchery near Springfield. There the heated discharge from a paper mill warms the icy water of the McKenzie River in the winter to a temperature of 11 degrees C. (52 degrees F.) for incubating coho eggs and rearing the fry. When the fry reach the smolt stage in the late spring, they are trucked to the coast at Newport and put in ponds that receive salt water pumped from the ocean. After from 10 to 14 days in the ponds the fish become adapted to full-strength ocean water. After a final feeding they are released into the ocean. In about 18 months they home on the artificial salt-water river that flows out of the ponds from which they were released, delivering themselves in prime condition for the harvest. Boats and nets are not needed. It is not even necessary for the ranchers to round up their animals and herd them into a corral; the fish go into the ponds on their own.

In the wild the pink salmon and the chum are usually in the more northerly parts of the salmon range. Ordinarily they spawn close to the sea; some races even spawn on tide flats, where fresh water seeps into the sea. After incubating in the gravel during the cold winter months the young fry pop out of the nest in the spring and go almost immediately into the sea. Unlike the other salmonids, the pink and the chum salmon can smolt as soon as the yolk sac is absorbed.

Most students of salmon biology have believed the chum and particularly the pink salmon could not live in fresh water but had to go to sea to survive. It was therefore surprising to find that a small release of pinks into Lake Superior showed up as adults three years later. This odd-year return is unknown in sea-going pinks, which normally return after two years. Undoubtedly a fishery will develop in the Great Lakes to exploit this new stock.

In the years after World War II huge hatchery programs were built up in northern Japan and on the island of Sakhalin in the U.S.S.R. to enhance the production of pink and chum salmon. Nearly all the returning salmon are caught live in traps or set nets rather than from boats as they are in most of the North American salmon fisheries. They are therefore unmarked and in prime condition when they are delivered to market.

The well-coordinated Japanese program has expanded so that more than a billion chum fry are being released each spring. The young smolts leave the coasts of Japan for ocean feeding grounds south of the Aleutian Islands and in the Gulf of Alaska. After from three to five years in the ocean they return to the Japanese rivers from which they migrated to sea. In the fall of 1982 the harvested return was well over 20 million salmon, compared with annual catches of from two to three million soon after the turn of the century, when the fishery was based on naturally spawning populations.

The Russians are releasing millions of pink and chum fry from hatcheries into the small streams of Sakhalin. They plan to increase the number by 100 million per year until the total reaches three billion in the year 2000. Like the Japanese, the Russians favor terminal harvesting. The spawners swim up collecting troughs along the riverbanks. The fishermen standing along the banks can pick them up and drop them directly into the processing lines. Terminal harvesting is by far the most economical and efficient way to harvest such a resource. It also makes possible accurate counts and thereby encourages precise and effective management.

The Atlantic salmon has been considered by some experts to be an endangered species, but a brighter future is beginning to emerge. National policy in Iceland has succeeded in maintaining and even increasing the number of salmon returning to Icelandic streams. The policy aims at protecting the resource rather than accommodating the commercial fishing industry as unfortunately many governments in Europe and North America do.

The salmon returning to Iceland belong to the farmers who own the land along the rivers. Only in the lower reaches of four of the 250 rivers and at three locations along the coast, however, are the farmers allowed to follow the ancient practice of netting salmon. This net fishery brings in about 40 percent of the annual catch. The rest is harvested by sport fishermen using rod and line. Both commercial and sport fishing rights must be leased from the farmers who own the land. The management of the salmon fisheries is strictly controlled by the government, which sees to it that an ample number of fish are allowed to proceed to the spawning grounds.

That program has been supplemented by artificial rearing and by the opening of rivers formerly impassable to salmon. At the government's experimental fish farm in Köllafjördur much progress has been made in the use of geothermally heated water for the accelerated rearing of Atlantic-salmon smolts, in controlling the length of time the fry are exposed to light each day to promote feeding and growth, in developing brood stocks for seeding in streams newly made accessible to salmon and in mixing salt water in rearing ponds to trigger the smolting process. Streams formerly impassable to salmon have been opened by building fish-passage facilities. The government also promotes and aids the development of commercial salmon ranching.

In western Europe the runs of salmon have been greatly reduced from the level of preindustrial times, not only by the unintentional impairment of habitats but also by interceptions at sea in which the drift nets of one country take salmon that originated in another country. In Scotland, Ireland, Norway and Canada entrepreneurs are taking steps to ensure the supply of Atlantic salmon for their markets by rearing the fish to maturity in sea ponds and floating cages. Norway, with its long coastline of protected fjords, is the leader in this new type of sea farming. In 1981 Norway marketed 8,000 tons of farmed Atlantic salmon; the output is expected to triple over the next three years.

Of all the salmonids the rainbow trout exhibit the greatest genetic diversity. From their original range along the margin of the eastern Pacific they have spread or been spread to almost every part of the globe. They thrive even in the Tropics at higher elevations.

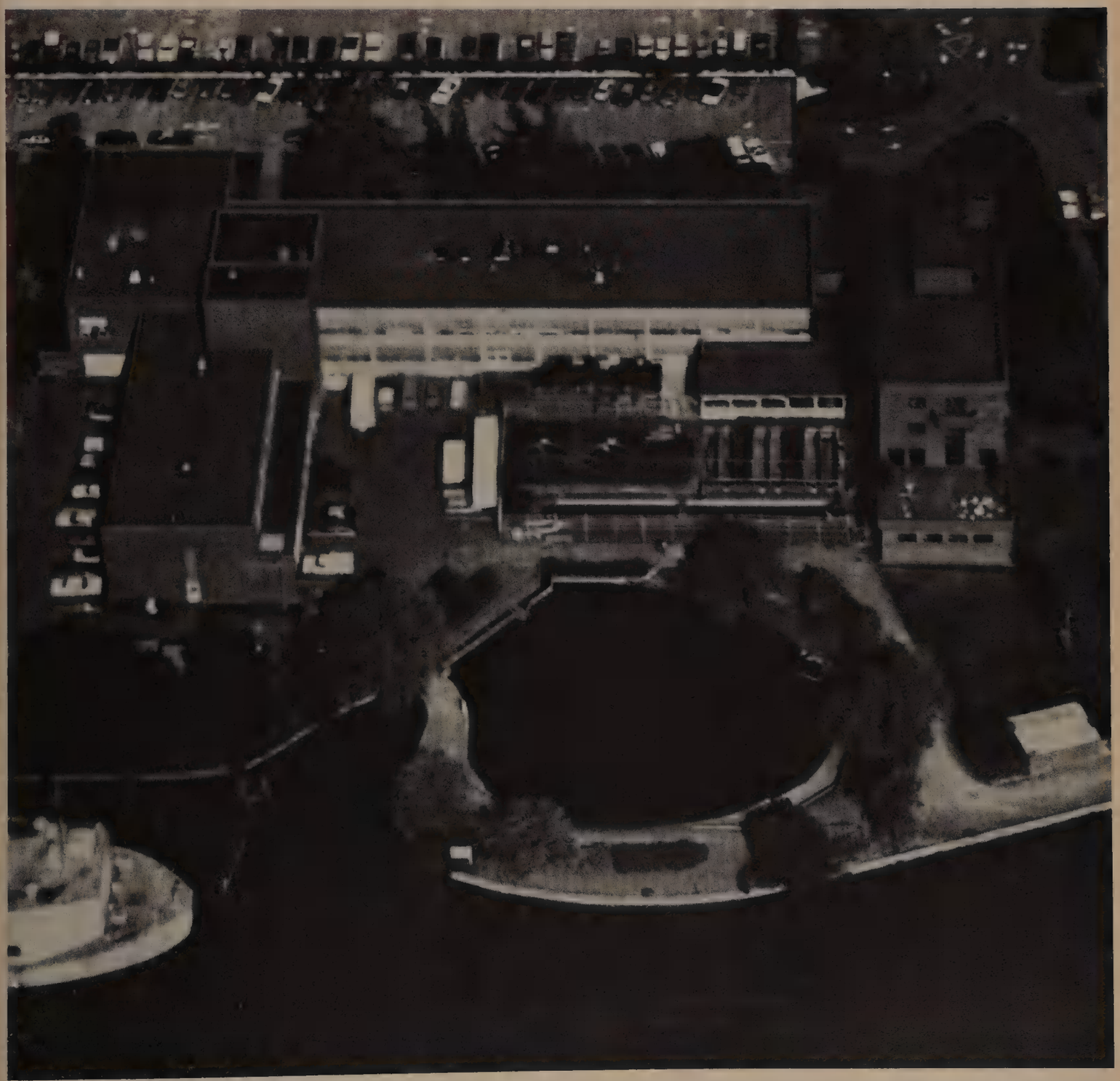
In addition to providing valuable

sport fisheries the rainbow trout are the basis of big commercial enterprises. Italy alone produces 21,000 tons of rainbow trout per year for marketing; Denmark, Finland, France, Norway and the U.K. all have major commercial operations. In southern Idaho, in the valley of the Snake River, trout farms yield about 13,000 tons of rainbow trout per year.

In the wild state rainbow trout spawn only in the spring. Now selectively bred stocks are available that spawn in nearly every month, providing a sound basis for commercial trout farming.

A program of selective breeding of rainbow trout has been in operation at the College of Ocean and Fishery Sciences of the University of Washington since 1932. The objective has been to turn out large numbers of healthy, acclimated fish for lake- and stream-management programs and for commercial farms. In the early years the trout reached maturity in their fourth year, weighing on the average 680 grams (1.5 pounds). Females yielded from 400 to 500 eggs in the first spawning. Some males of our select stock now reach 680 grams in their first year. The females, which mature in their second year at an average weight of 4,500 grams (10 pounds), yield about 10,000 eggs each. Although the rate of increase in the average length of the fish selected as brood stock has slowed, the average weight of the fish continues to increase. The result is a deep-bodied fish with better flesh. We like to think our results compare with the poultrymen's development of the broad-breasted turkey.

The university's strain of rainbow trout has been cut off from the sea for many generations. Nevertheless, and somewhat to our surprise, smolts from eggs of this strain adapted well when they were introduced to floating sea cages in Norway. They grew to a large size (from five to 15 pounds) in what



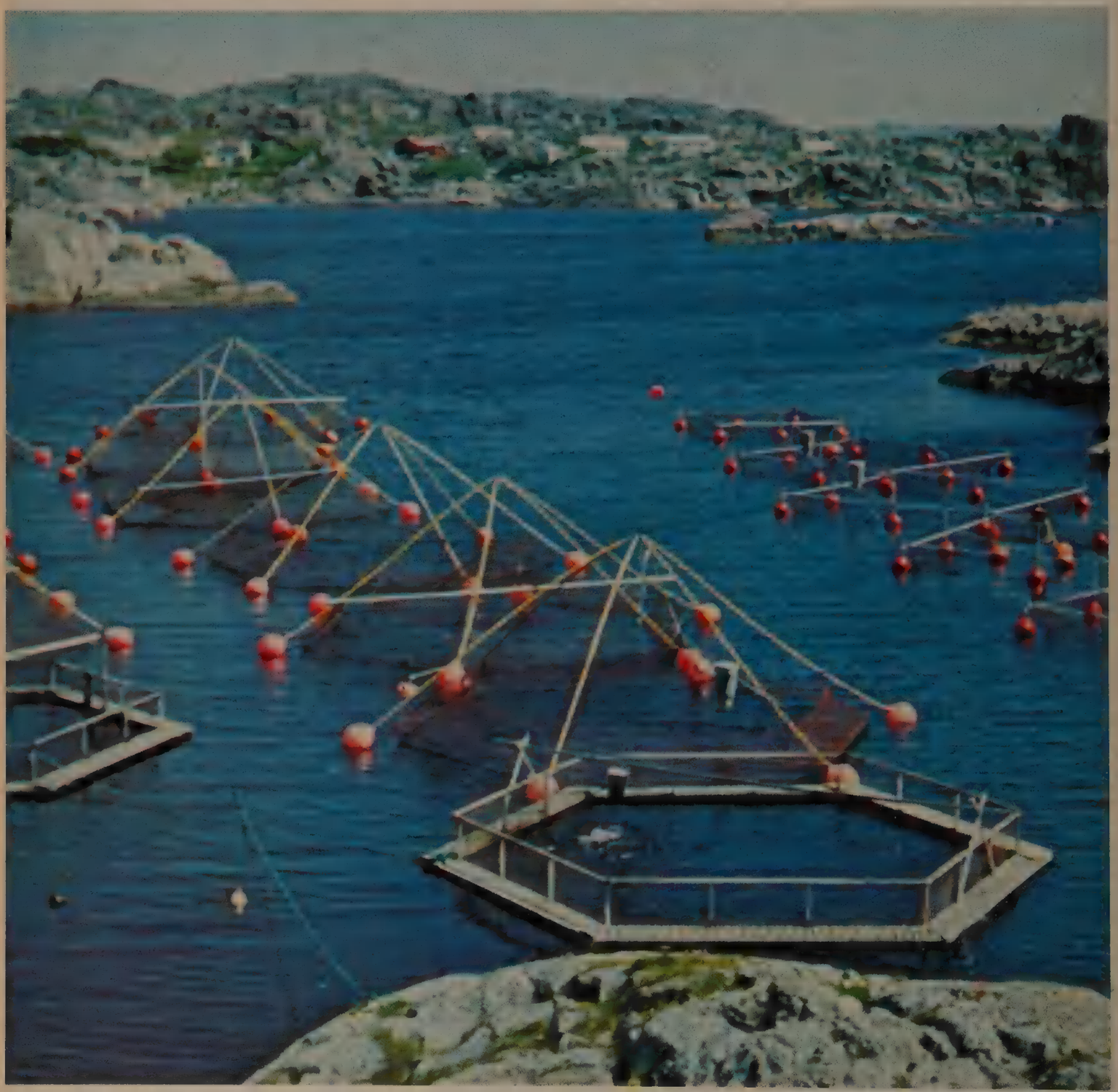
"HOME STREAM" for the chinook and coho salmon reared artificially at the University of Washington is a manmade pond at the edge of the campus. To reach it the 5,000 to 6,000 salmon that return to the pond from the sea each fall have to make several turns and pass through a ship canal. The water that attracts them is pumped from the canal to the pond and flows back to the canal by gravity.

we thought would be a quite stressful environment.

The rearing of rainbow trout in sea cages has become a profitable business in Norway. Some 4,000 tons were marketed in 1981, mostly in Europe. The market is expanding to the U.S., where large, fresh, pen-reared rainbows from Norway now appear in Seattle, the salmon capital of the world.

The crossbreeding of closely related stocks or races of plants and animals often results in "hybrid vigor." This tendency can be observed among the salmonids. Interracial crosses between steelheads (anadromous rainbow trout) and the select nonmigratory rainbows developed at the University of Washington result in hybrids that grow faster than the steelheads but slower than the nonmigratory rainbows. It appears that the dominant growth factor in the university stock has been transmitted, at least in part, to the steelheads. Moreover, the hybrids seem to retain the steelhead characteristic of migrating to the sea, where they grow rapidly. When they return at maturity, they provide exciting sport for recreational fishermen because they have the rainbow's trait of biting at a lure.

The selective breeding of salmonids can in a relatively short time affect their form, size, behavior, taste and food value as profoundly as 1,500 years of stock breeding have affected dairy and beef cattle. The future evolution of salmonid species now lies chiefly in the hands of the people who are beginning to breed them on a large scale. The task embodies challenges. One challenge is to maintain genetic diversity by providing appropriate sanctuaries to ensure the survival of wild stocks. Another is to utilize the enormous adaptability of these remarkable animals by selectively breeding stocks that will take hold both in new environments and in places where native stocks have been displaced by human activities.



FISH FARMING of Atlantic salmon and rainbow trout is carried out by Norwegian entrepreneurs in floating cages of the kind shown here. In 1981 Norway marketed more than 8,000 tons of farmed Atlantic salmon (*Salmo salar*). The output is expected to triple by 1986.

1700 pounds of optical perfection.

Perfection is not a term we use lightly. Or frequently. We've been taught for generations that perfection, in lens-optics terms, means ¼ wavelength. We've bettered this tolerance in the manufacture of an "egg crate" mirror, one-fifth the weight of a solid mirror. Manufactured for the Space Telescope Program, every point on its nearly 8-foot-diameter aspheric surface is within ⅒ wavelength of perfection. So while we're not 100% perfect, we're within several millionths of an inch of that goal.

We put all our experience into play in the manufacture of this massive optical element 97½ inches in diameter and 1700 pounds in weight. And our challenges were many. Since this lens is intended to perform in orbit, where the pull of gravity has been neutralized, we had to solve a gravity-distortion problem during the manufacturing cycle. We solved the problem with an air cushion, a membrane that inflates for testing to provide a uniform support and thus eliminate gravity-induced strain. The support deflates for polishing operations.

An aspheric mirror to gather information from the universe is an example of our large-optics capability. (Optics of this kind are ground, polished, and tested in our own lab with its vibration-isolated, six-story-high interferometric test towers.) But we're big in small things, too. Like lenses for cameras and copiers. We meet customers' needs with pride, precision, and stringent quality control.

In addition to massive, sophisticated equipment, a century of optical progress, and systems utilizing our ability to the fullest, we offer customers a tradition of realism when predicting costs and delivery.

That's important. Whether you seek to get a clearer picture of your neighbor or of celestial objects.

For more information about Kodak's contribution to the Space Telescope Program, write for the booklet *Mirror for the World's Most Powerful Telescope.* Send your request to: Richard M. Price, Research and Engineering, Kodak Apparatus Division, 901 Elmgrove Road, Rochester, NY 14650.

Kodak

Kodak. Where technology anticipates need.



© Eastman Kodak Company, 1983

SCIENCE AND THE CITIZEN

Out to Launch

For more than two decades it has been the official policy of the U.S. to structure the nation's strategic nuclear forces in such a way as to ensure their ability to "ride out" a full-scale attack by the U.S.S.R. before the order is given to launch a retaliatory attack. The delayed-response strategy is intended to minimize the chance of starting a nuclear war by accident, on the basis of an electronic false alarm, say, or perhaps a human error. At the same time U.S. military and political leaders have been careful to avoid ruling out in public the possibility that a retaliatory second strike might be launched sooner, either on receiving early warning that a Russian first strike was on the way or just after it was learned that the incoming warheads had begun to reach their targets. The rationale offered for leaving open such a "launch on warning" or "launch under attack" option is that it adds to the uncertainty of military planners in the U.S.S.R. and hence reinforces the credibility of the U.S. deterrent.

Recent statements by members of the Reagan Administration have tended to draw increased attention to the early-launching option. The issue has been raised in the context of efforts by the Administration to get congressional approval of its plan to deploy 100 MX missiles in existing Minuteman missile silos. Advocates of the MX program had previously argued that the same silos were in danger of being destroyed by increasingly accurate Russian missiles; indeed, the vulnerability of the Minuteman system was advanced as one of the main justifications for proceeding with the MX program. Opponents of the Administration's newest basing scheme for the MX have pointed out that in spite of plans to "superharden" the existing silos, MX missiles based in them would be no more likely to survive a Russian "counterforce" attack than the Minuteman missiles they were designed to replace. Indeed, because the huge, 10-warhead MX missiles would constitute a more attractive target than the smaller three-warhead Minuteman missiles, the end result of this part of the Administration's "strategic modernization" program could well be a less survivable land-based missile force.

In testifying before the Senate Appropriations Committee in April, Secretary of Defense Caspar W. Weinberger and Chairman of the Joint Chiefs of Staff General John W. Vessey, Jr., acknowledged the theoretical vulnerability of MX missiles deployed in Minuteman silos. According to General Vessey, no more than 30 percent of the missiles could be expected to survive a Russian counterforce attack "if we ride out the attack." He added, however, that the Russians "have no assurance that we will ride out the attack." When the question was raised of whether the Administration was therefore moving in the direction of a strategy of launching under attack, Secretary Weinberger refused to discuss the matter further in public.

Although the statements of officials such as Secretary Weinberger and General Vessey on the release of nuclear weapons do not differ fundamentally from those of their counterparts in previous administrations, they do seem to reflect growing pressures on the leaders of both the U.S. and the U.S.S.R. to resort to the early-launching option in formulating their nuclear-war plans. Another indication of the concerns engendered by the advent of a new generation of "counterforce-capable" weapons in the arsenals of the two superpowers is the renewed attention being paid to command, control and communications (C^3) systems, which are generally thought to be the most vulnerable part of any deterrent force. The comparative "softness" of C^3 targets means they are vulnerable in theory not only to highly accurate long-range intercontinental ballistic missiles (ICBM's), which would take about 30 minutes to reach their targets, but also to shorter-range submarine-based and "theater" nuclear weapons, which have flight times of 15 minutes or less. Future discussions of arms-control issues are likely to focus increasingly on the dangers inherent in the seemingly inexorable technological trend toward reduced decision-making time and on the prospects for improving the reliability of systems designed to prevent the outbreak of an accidental nuclear war in the emerging "hair trigger" environment.

Reactor Power

Six countries now generate more than 25 percent of their electricity with nuclear power. (The U.S. has 72 commercial reactors in operation, far more than any other nation, but because of the country's size they account for only 13 percent of the power.) In France 32 operating reactors produce more than 40 percent of the power; France intends to have 85 percent of its installed capacity nuclear by the end of the century, with essentially only peak loads supplied by other sources. Nuclear reactors are generating power for the electric grid in 23 nations besides the U.S., and 17 more nations have firm programs for nuclear power. By the end of last year 9 percent of the world's electricity was being generated by nuclear reactors, up from 8 percent in 1981.

The data on the 40 nations other than the U.S. with nuclear-power programs were assembled by the Atomic Industrial Forum, Inc. In terms of reliance on nuclear energy the survey found the leaders other than France to be Sweden (36 percent), Finland (35.8), Switzerland (28.1) and Belgium (25.4). In terms of number of operating reactors the leaders were the U.S.S.R. (37, with no figure available for how much of the country's electricity they generate), France and the U.K. (32 each), Japan (25), West Germany (15) and Canada (12).

The survey tallied 207 operable reactors outside the U.S., eight more than there were at the end of 1981. In other categories there were 163 reactors under construction, 13 on order and 172 planned. The total capacity of the operating plants is 105,823 net megawatts, up 9 percent over the figure for 1981.

The survey took note of four "striking developments overseas" in 1982. One was the first order for commercial nuclear reactors from the People's Republic of China; it calls for two 300-megawatt pressurized-water reactors to be built near Shanghai by 1988. Another was the entry of Brazil into the ranks of the countries with operating commercial reactors. The third was "the start of full commercial operation by both India and Japan of spent-fuel reprocessing plants, signaling a major move toward nuclear self-sufficiency for both nations." The fourth was that South Korea became the eighth country to have a program for fast-breeder reactors.

Four in Every Ten

Forty percent of the world's people live in India and China. Therefore government policies bearing on population growth in the two countries have a significant relevance to the size and well-being of the human population as a whole. The recent demographic histories of India and China offer an instructive comparison, in particular because of the notable similarities between the two countries. Both are relatively poor, heavily agricultural nations. Both had a largely colonial status until after World War II, when they gained economic independence through powerful revolutionary movements. Reducing the birth rate has been a goal held strongly by government officials in both countries. There is, of course, a striking political and economic difference between India and China. Whereas the economy of In-

dia is largely a market one, the economy of China is almost entirely state-controlled. In a recent issue of *Proceedings of the National Academy of Sciences* Ansley J. Coale of Princeton University summarizes the significant demographic disparities that have resulted partly from the political differences between the two countries.

The current population of China is one billion. The population of India is 700 million. Between 1951 and 1981 the Chinese population increased by about 80 percent and the Indian population by about 90 percent. In China, however, there was a substantial reduction in the overall rate of population growth in the 1970's. There was no corresponding reduction in India. The overall rate of population growth is determined by the birth rate and the death rate. In both countries the death rate after World War II was high. In China in the 1950's there were 18 deaths per 1,000 people per year. By the 1970's the rate had been reduced to seven per 1,000, roughly equal to the rate in many developed countries. In India, on the other hand, the death rate in the 1950's was 26 per 1,000 and by the 1970's it had decreased only to 17 per 1,000.

According to Coale, the greater improvement in mortality in China is due largely to the fact that the system of medical care there is better and more egalitarian than the one in India. China has a higher ratio of physicians to population than India. The Chinese physicians are distributed more equitably than they are in India, where the doctor-patient ratio is eight times higher in the cities than it is in the countryside. Moreover, China has an extensive government-sponsored system of paramedical personnel, many of whom practice in rural areas. In both India and China more than 80 percent of the people live in the countryside. The better health care received by the rural majority in China appears to account for much of the difference in mortality rates.

In China fertility has decreased along with mortality. The birth rate in the 1960's was 36 per 1,000 of population. In the 1970's the rate was 23. The decline was due mainly to a vigorous official family-planning program. The intense coordination of all social and political activity in China contributed much to the effectiveness of the program. Such all-embracing social coordination raises the possibility that coercion is practiced in family planning. Chinese couples must obtain a "planned-birth certificate" before the wife becomes pregnant. Coale notes that there have been reports of forced abortions in instances where a woman was pregnant without a permit.

Coercion has also been a significant issue in India, where official family-plan-



The 757 is the most fuel-efficient jetliner in the sky. In passenger comfort, it's superior to any other plane its size. Here's the perfect match of advanced technology and passenger conveniences. It's one way Boeing is helping to keep air fares one of the world's best travel values.

BOEING
Getting people together.

In 1982, Porsche won the World Championship of Makes. Audi won the World Rally Championship for Manufacturers and SCCA Pro Rally Championship.

What we learn from our competition cars, we put into our production cars.

And now you can arrange to drive your new Porsche or new Audi where it was built, tested, and refined—with our Delivery In Europe Program.

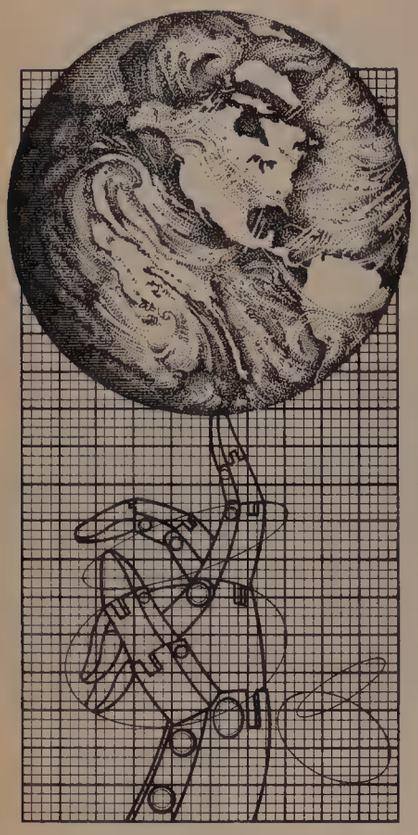
Simply visit your local Porsche Audi dealer and order the model of your choice: the Porsche 928S, 911SC, or 944. Or the new Audi 5000S, the 4000S, Coupe GT, or all-wheel drive Quattro. Then take delivery in Europe. And after your stay, we'll deliver your car to a U.S. port free of charge.

For more information, see your Porsche Audi dealer: the home of champions.

Visit Europe And Drive Home A World Champion.



Future Pod '83



It's a space in time, a look at the future. It's a trip through the imagination and just one of the many attractions to be experienced at Ontario Place, Toronto. And it's free with your admission to the grounds. Come touch, see, feel and live our world of tomorrow, today. Talking robots, computer-controlled arms, energy alternatives and the shape of things to come in transportation and communications. See nuclear power at work, laser imagery and cars of the future. Sooner or later you'll experience the future.

For more information about Ontario Place and Future Pod '83 call, TOLL FREE; **1-800-828-8585**; in New York State **1-800-462-8404.**

ONTARIO yours to discover!

ning efforts have been much less successful. The birth rate in India in the 1970's was 39 per 1,000 people per year, only very slightly less than the rate in the 1960's. According to Coale, "despite official support (for 30 years) of family planning, the government of India has not been able to organize a birth-control program that regularly provides adequately staffed services for most of the population." Partly out of frustration at past failures, the government of Prime Minister Indira Gandhi instituted a draconian sterilization program under the Emergency Act of 1975. Charges that Indian villagers had been sterilized against their will were soon widespread. Coale notes that "coercion was denied by the government but was almost inevitable, whether or not intended, because, for example, local officials set quotas for villages and for the number of motivated couples a schoolteacher was to round up." Resentment of the sterilization program contributed to the defeat of the Gandhi government in the 1977 elections. Thereafter family planning disappeared from the national agenda, to return only with Mrs. Gandhi's reelection.

Clearly both the Chinese and the Indian government are committed to reducing the rate of population growth. In explaining the much greater success of the Chinese in actually doing so, Coale cites their form of government and also the Chinese culture, which could be more conducive to the use of contraceptives and to the idea of limiting family size than Indian culture is. In spite of the past successes of the state birth-control program, however, the Chinese population will continue to increase for several decades because of the large number of young people who are now reaching childbearing age. Coale estimates that the population could cease to grow in 2020, when there will be some 1.2 billion Chinese. The smaller Indian population has a much higher rate of growth. Coale estimates that if fertility in India begins to fall soon, the ultimate stable size of the Indian population could also be about 1.2 billion. He concludes that in both countries "a lower birth rate now is desirable, but the ideal rate is not zero. There are social and political costs of excessive emphasis on the immediate achievement of very small families; the rights and sensibilities of the current population and the disequilibrating effects of drastic changes in age composition must enter the calculation of desirable population policies."

Still at Large

Recent experimental successes in elementary-particle physics have been confined largely to detecting particles that can exist only in the extreme environment of the high-energy accelerator. Two of the three intermediate vector bosons, the W^+ and the W^-, are inferred from the products of five energetic collisions seen earlier this year by two groups of investigators at the European Organization for Nuclear Research (CERN). As of this writing four more events have been recorded, bringing the total to nine. The third intermediate vector boson, the Z^0, may have been detected as well; a single pair of electrons, each with an energy of about 50 GeV (50 billion electron volts), were observed moving in opposite directions from the site of a collision between a proton and an antiproton. Such an event is widely regarded as a good signature for the presence of the Z^0.

A preliminary account of an analysis of the intermediate-vector-boson data, done by theorists at CERN, also suggested that the long-sought *t*, or top, quark has been found; more careful analysis must be carried out, however, before the existence of either the Z^0 boson or the *t* quark can be confirmed. Mesons that show naked beauty, or have a quark composition for which the property *b*, or beauty, does not cancel, have been observed following high-energy collisions between electrons and positrons (see "Particles with Naked Beauty," by Nariman B. Mistry, Ronald A. Poling and Edward H. Thorndike, page 106). The detection of all these exotic particles is important for the confirmation of theory, but in other respects the particles are ephemeral: they may not exist anywhere in the universe outside the instruments of physicists.

There are several phenomena, however, that remain undetected but should theoretically be found in ordinary matter. They are among the most sought-after and most elusive prizes in physics. Three fundamental questions have been raised: Is there a magnetic monopole? Does the neutrino have mass? Does the proton decay? Not only would affirmative answers to such questions be spectacular confirmation of theoretical predictions but also any definitive answer would have profound implications for efforts to forecast the long-range future of the universe. Nevertheless, in spite of years of work, no definitive answers have emerged; experimental physicists have only continued to narrow the limits within which the phenomena can be detected.

The monopole has attracted attention because its existence is predicted by the grand-unification theories, which link the strong, the weak and the electromagnetic force, three of the four fundamental forces of nature. (The fourth is gravity.) According to such theories, isolated magnetic poles having a mass of about 10^{16} times the mass of the proton should have been created within the first 10^{-35} second of the big bang. Such particles

are many orders of magnitude more massive than the most massive particles that can be generated in accelerators; nevertheless, if they were created at the beginning of the big bang, most of them should still exist, perhaps as dense matter collected at the core of the earth or as micrometeoroids gravitationally bound to our galaxy or to the solar system [see "Superheavy Magnetic Monopoles," by Richard A. Carrigan, Jr., and W. Peter Trower; SCIENTIFIC AMERICAN, April, 1982].

The monopoles of cosmic origin should be the easiest to detect. Blas Cabrera of Stanford University recently devised an instrument in which three coils of superconducting niobium wire are mounted along three mutually perpendicular axes. If a monopole were to pass through the coils or through a shield that is mounted outside the coils, a small but measurable current would be induced in at least one of them. Cabrera found evidence for a single monopole last year by detecting the change in current in a smaller, less elaborate coil, but it now appears likely that the signal was spurious. With the new device Cabrera has seen nothing; if the mass of the monopole is indeed 10^{16} times that of the proton, his data now indicate there can be no more than one monopole for every 10^{17} protons in the universe.

Cabrera's superconducting coils can detect monopoles moving at any velocity, but the sampling rate is low because the coils must be relatively small. A scintillation counter can sample matter more rapidly, but slow-moving monopoles may evade its detection. Investigators from Stanford and from the University of Utah have attempted to find monopoles with a scintillation counter. They report that no monopoles are observed having velocities between 1.4×10^{-4} and 3×10^{-2} times the speed of light. The intensity of monopole radiation in the vicinity of the earth is probably less than 5×10^{-12} per square centimeter per steradian per second.

If supermassive magnetic monopoles are found, they could make up a significant part of the total mass of the universe. The attempt to detect a nonzero mass for the neutrino is also partly motivated by cosmology: if neutrinos have mass, the sum of it may be sufficient for gravity to arrest the expansion of the universe and drag all matter and energy back into the primordial singular point from which it began.

One mass-detection strategy takes advantage of the fact that there are three kinds of neutrino; if neutrinos have mass, quantum mechanics implies the three kinds of neutrino should oscillate, or freely exchange their identities. Investigators from the California Institute of Technology, the Technical Universi-



Passengers rate the Boeing 767 as a superb flying experience. It has two wide aisles and is the only wide-body with seven-abreast seating, so nearly everyone has either a window or aisle seat. Pilots rate the 21st Century flight control system as the most advanced in the world. BOEING
Getting people together.

A BERLITZ® LANGUAGE COURSE
in a handsome briefcase
for only $125
if *you* pay for it
BUT
. . . wouldn't your *company*
be interested in having you
speak a second language?



These days a second language can help you help your company and so help *you* earn more. The many hispanic nationals now in our work force respond more readily to someone who speaks their language. Also, selling today is international. Whatever language it would pay you to speak, Berlitz has a do-it-yourself course in a handsome briefcase you can take anywhere. Contents—90 minute introductory tape, 40 lessons on five 60-minute cassettes, 6 illustrated word and phrase books, rotary verb finder to help you find verb tenses instantly. If not satisfied, return within 10 days and get your money back. Learn from the leader!

For your convenience on credit card orders dial toll-free
1-800-228-2028 ext. 35
and **refer to Dept.3112**
24 hours a day,
7 days a week.

Why not give us a ring—right now! (our expense)

(Nebraska residents dial 402-571-4900)

Berlitz Publications, Inc., Dept.3112
3490 Lawson Blvd., Oceanside, NY 11572

Send **Berlitz Comprehensive Cassette Course(s)** checked. $125 each in briefcase plus $4 for shipping and insured delivery.

☐ **French** 86100 ☐ **German** 86101 ☐ **Italian** 86102 ☐ **Spanish** 86103

Enclosed check _____ money order _____ payable to Berlitz.
Or charge my AMEX _____ Diners Club _____ VISA _____
MasterCard Interbank # _____ Exp. Date _____
Card # _____ Exp. Date _____
Name _____
Address _____
City _____ State _____ Zip _____

N.Y. residents add sales tax. Allow 4 weeks for delivery.

SPACE CALENDAR
JULY 4-10, 1983

Just Slightly Ahead of the Times

Published Weekly, Mondays
by
SPACE AGE REVIEW INC
SPACE CALENDAR ™
The TV Guide™ of Space Development
tells you
★ What's Up
★ Who's Who ★ How $$$ Much
★ How and Why ★ When, Where
about
★ Launches
★ Space Technology Developments
★ NASA and STS Activities
★ Private Space Entrepreneurs
★ Space Conferences and Meetings
★ International Developments
Theme for July 4-10, 1983 Issue
Space and America

Professionals, *leaders, career people planning to travel, work and live in space need to know what's up,* ***subscribe:*** *Rates $69/year Individual; $39 Introductory First Year Special for Members of Space Development Organizations AAS, AIAA; L5 (please include membership #); $99/year Institution. Add $30 Overseas (please remit in US currency). $3 for preview copies.*

SPACE CALENDAR ™
3210 Scott Blvd—Dept SCA
Santa Clara, California 95051
408-988-0592

SCIENTIFIC AMERICAN

is now available to the blind and physically handicapped on cassette tapes.

All inquiries should be made directly to RECORDED PERIODICALS, Division of Volunteer Services for the Blind, 919 Walnut Street, 8th Floor, Philadelphia, PA 19107.

ONLY the blind or handicapped should apply for this service. There is a nominal charge.

ty of Munich and the Swiss Institute for Nuclear Research are measuring the flux of neutrinos outside a nuclear power reactor at Gosgen in Switzerland. The apparatus is sensitive to the presence of the electron neutrino; if electron neutrinos change into muon neutrinos or tau neutrinos as they leave the reactor, the population of electron neutrinos should decrease with distance from the reactor. Felix H. Boehm of Cal Tech reports that no such variations are observed. Therefore neutrino oscillations, if they exist, are quite small.

The oscillation experiments do not give direct upper limits on the mass of the neutrino; they only make it possible to estimate the square of the difference between the masses of two kinds of neutrino. An experiment done three years ago by Russian investigators estimated the neutrino mass directly by measuring the decay rate of tritium, the radioactive isotope of hydrogen. They concluded from their results that the neutrino has a small but finite mass, between 14 and 46 electron volts. Reanalysis of the Russian data by J. J. Simpson of the University of Guelph, however, suggests that the data are consistent with a zero mass for the neutrino. Boehm's group has recently set an upper limit for the neutrino mass of 10 to 20 electron volts; the group measured the rate of the decay process called double beta decay in the isotope germanium 76. A larger-scale experiment is now planned at Cal Tech, where double-beta-decay rates will be measured in 3,000 liters of xenon 136.

Perhaps the most sobering experimental result for grand-unification theories is a new lower bound on the lifetime of the proton. In the simplest form of grand-unification theory the proton should decay in less than 10^{29} years, with the large possible error of plus or minus 2 in the exponent. Such infrequent events can be detected by searching for the proton decay in a large mass of ordinary matter. Investigators from the University of California at Irvine, the University of Michigan and the Brookhaven National Laboratory are monitoring 8,000 tons of water, or about 10^{33} protons, in a salt mine near Cleveland, but to date no event that resembles a proton decay has been observed. The findings imply the lifetime of the proton is greater than 6.5×10^{31} years, which is too long for the prediction derived from the simple form of grand-unification theory even if the potential error in the prediction is taken into account.

There still remains some comfort for the theorists. Two other experiments have reported evidence that the proton decays. In the mines at Kolar Gold Fields in southern India three candidate decay events have been reported. Investigators at CERN, monitoring plates of iron suspended in the Mont Blanc Tunnel between France and Italy, have announced one candidate event. The jury remains out on whether or not these results are real proton decays, and if they are, how such results can be reconciled with the null result in Cleveland. In any case, according to Maurice Goldhaber of Brookhaven, "the idea of grand unification will live forever." There are many more complex versions of the attempt to unify the forces of nature that may also get their day in court.

Hot Potato

The cultivated potato plant, *Solanum tuberosum,* is susceptible to a variety of insect pests. The wild potato *Solanum berthaultii* is not. Its resistance seems to be associated with two kinds of tiny glandular hairs that cover its stems and leaves; aphids in particular have been thought to be trapped physically by the hairs and their sticky exudates. The antiaphid strategy of *S. berthaultii* turns out to be more sophisticated than that. Investigators at the Rothamsted Experimental Station in England have found that one type of hair releases a hydrocarbon that is the main component of an aphid pheromone. A pheromone is a hormonelike substance that is secreted by an animal and influences the behavior of other individuals of the same species, and this particular pheromone serves as an alarm signal among aphids. In effect the plant masquerades chemically as an aphid warning its fellow aphids to keep away.

R. W. Gibson and J. A. Pickett report in *Nature* how they found the bogus pheromone by extracting *S. berthaultii* leaves with alcohol and analyzing the extract by gas chromatography and mass spectrometry. Of several hydrocarbons present in the extract but not in a similar extract of cultivated-potato leaves, one was shown to be (*E*)-beta-farnesene, the aphid pheromone. Its source was found to be a sticky droplet at the tip of the type-*B* hairs. To demonstrate that the substance is volatile enough to affect approaching aphids, Gibson and Pickett collected air from above intact leaves, passed the air through a solvent, pentane, and measured the amount of farnesene that had been dissolved in the pentane. Some 50 billionths of a gram of the pheromone was thus extracted from 20 milliliters of air, a concentration high enough to generate an alarm response.

The Rothamsted workers went on to demonstrate such a response directly. *S. berthaultii* leaves were put in a syringe and air was expelled from the syringe toward a colony of wingless aphids (*Myzus persicae*) about a centimeter away. The aphids dispersed, showing the same

behavior as aphids exposed to authenticated samples of the pheromone; air from an empty syringe or one loaded with leaves of the cultivated potato had no such effect. Then aphids were placed a centimeter away from a potato-plant leaf, facing toward the leaf. Most of the aphids walked right onto the leaf of a cultivated potato plant. Only six of 48 aphids tested did so in the case of a wild-potato leaf; the others either turned away at a distance of from one millimeter to three millimeters or walked parallel to the edge of the leaf, maintaining a similar distance. Again the behavior is like that of aphids faced with a plant treated with (*E*)-beta-farnesene.

It is not hard to cross *S. berthaultii* with the cultivated potato, Gibson and Pickett point out. Some hybrid plants have the type-*B* hairs in quantity. If the ability to secrete the pheromone could be widely introduced into *S. tuberosum,* it should protect potato crops against the sapsucking damage done by aphids and also against viruses spread by the insects, a still more important threat.

Ice Hole

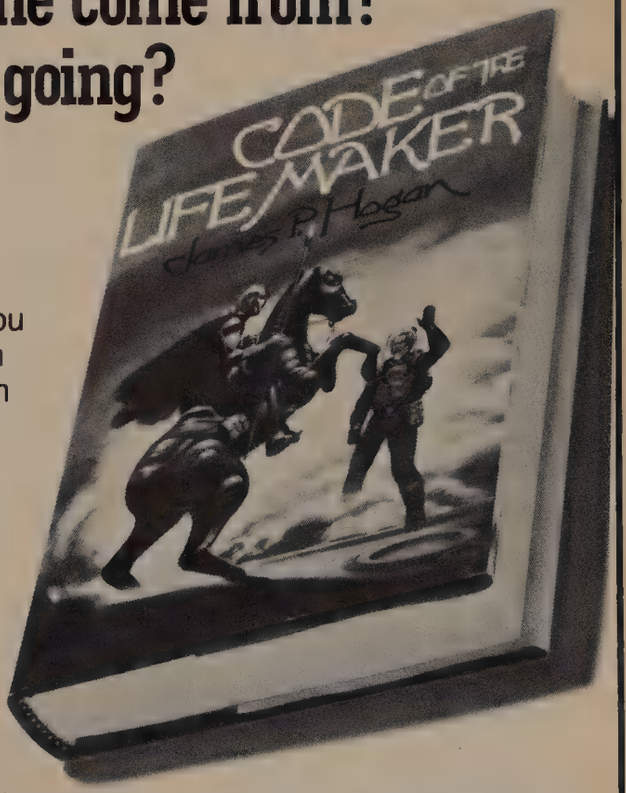
Few places on the surface of the earth embody the subtle, complex interactions that command the attention of earth scientists better than the ones where ice, air and water meet, and at these places of meeting few things exemplify the complexity better than polynyas: regions in frozen seas that stay anomalously free of ice. The word is the Russian for open place in the ice.

Two groups of investigators have contributed papers on polynyas to an issue of *Journal of Geophysical Research* devoted entirely to "marginal ice zones." J. D. Schumacher and C. H. Pease of the National Oceanic and Atmospheric Administration in Seattle and K. Aagaard and R. B. Tripp of the University of Washington investigated the polynya in the Bering Sea off the southern coast of St. Lawrence Island. There currents flowing north toward the Arctic Ocean meet winds blowing south. Ice forms continuously but the winds drive it off. At locations near the polynya the workers measured ocean currents, temperature and salinity and also wind velocity from November, 1980, through June, 1981. About two-thirds of the significant fluctuations in the currents detected that winter could be accounted for by variability in the large-scale pattern of oceanic circulation or by "direct wind forcing." Eleven times, however, something more complex happened. The ocean current reversed, and the water temperature decreased, so that it remained near the freezing point of the salt-laden sea. Evidently "brine rejection" (the expulsion of salt from freezing ocean water) increased the density of



World travelers choose the 737 because it has one of the best on-time records. The 737-300 is the latest addition to the family. This new jetliner, with added passenger amenities, advanced flight controls and quieter engines, will make flying an even better travel experience.
BOEING
Getting people together.

What is man?
Where does he come from?
Where is he going?



In his extraordinary new novel, *Code of the Lifemaker,* James P. Hogan, the bestselling author of *Inherit the Stars* takes you to a future world in which mankind meets its match in an alien race of intelligent robots.

It's the most amazing novel yet from the man who puts the science back in science fiction.

Now on Sale
DEL REY A Del Rey Hardcover Published by Ballantine Books

The world's densest compute is now the heart of the world's mos 32-bit computer.

Our 32-bit CPU.

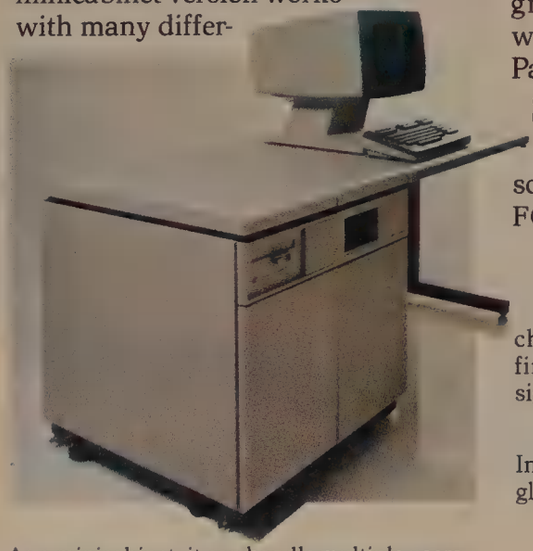
From time to time, miracles of technology come along to make previously impossible tasks not only possible, but easy. That tiny 450,000-transistor integrated circuit is one of those technological miracles.

Hewlett-Packard didn't develop it just to break the record for most transistors on a chip, but to put on an engineer's or scientist's desk a computer so powerful that it can do the work of mainframes costing four times as much.

32-bit computers for 32-bit applications.

The new HP 9000 computer based on this and four other 'superchips' can handle formidable engineering and scientific problems. The scientist solving complex systems of equations, the mechanical engineer doing finite element analysis or three-dimensional modeling, the electrical engineer analyzing complex circuits or designing very large-scale integrated circuits — these are the kinds of technical people and problems the HP 9000 family is designed for.

It comes in three versions. The integrated workstation is complete with keyboard, color or monochromatic graphics display, fixed and flexible disc drives, and printer. For systems manufacturers, there's a rack-mountable box. And for a variety of single-user and multi-user applications, the minicabinet version works with many different displays and peripherals.



As a minicabinet, it can handle multiple users.

A rack-mountable version is available, too.



All are true 32-bit computers, with 32-bit CPUs, memories, and data paths. And the multi-CPU architecture lets you nearly double or triple your processing power at any time by adding one or two CPU boards. Without increasing the computer's size.

Two operating systems are better than one.

The integrated workstation is available with a choice of operating systems. One is HP's highly evolved, high-performance Enhanced BASIC, augmented with 3-D graphics and a software innovation called a run-time compiler. This substantially increases program execution speed, while retaining an interactive development environment.

The other operating system, called HP-UX, is a fully supported, extended version of the popular UNIX® HP-UX, available on all HP 9000s, adds virtual memory, graphics, data base management, data communications, and enhanced file capability to the basic UNIX 'shell.' High-level programming languages available with HP-UX are FORTRAN 77, Pascal and C.

Software, and plenty of it.

Much of the vast range of existing software written in HP BASIC, FORTRAN 77, Pascal and C is transportable to the HP 9000. HP will also be offering proprietary software packages emphasizing computer-aided design and engineering. These will tie the HP 9000 into HP's Manufacturer's Productivity Network (MPN). Third-party software suppliers will be providing many of the most widely used CAE packages for 32-bit computer systems. And both HP 9000 operating



The 32-bit CPU chip is bonded to the finstrate which doubles as a signal carrier and heat sink.

Up to three CPU boards and three Input/Output Processors can fit into a single HP 9000.

TCG-207

hip
ffordable

systems offer extensive program development tools.

You also get a choice of communication tools. The HP 9000 is currently compatible with Ethernet,™ and with HP's Shared Resource Manager (SRM) which lets clusters of HP 9000 and 16-bit desktop computers share data and use common peripherals.

Links to central computers are also available. And in late 1983, HP will offer local area networks based on the IEEE-802 standard.

New technology from the silicon up.

The five superchips that make the HP 9000 possible are the 32-bit CPU, which can execute a million instructions per second; an eight-channel Input/Output processor (IOP); a random-access memory chip capable of storing 128K bits of data; a memory controller that 'heals' up to 32 bad memory locations; and an 18-megahertz clock.

Hewlett-Packard's advanced NMOS-III process makes it possible to put 450,000 transistors on a chip only 0.4 centimeters square. This tremendous density of electronic components could have required an expensive and elaborate cooling system.

Instead, HP engineers developed a new mounting structure called a finstrate, a copper-cored circuit board, which acts as both cooling fin and substrate. The finstrates containing the CPU, IOP, memory, and clock chips are housed in a lunchpail-sized module.

One user, one mainframe.

Clearly the trend in engineering and scientific computation is away from large machines shared by multiple users and towards networks of powerful personal workstations, sharing peripherals and data bases. The reason is compelling. An engineer or scientist in personal control of an HP 9000 can solve so many more problems more easily that the increased productivity alone makes the cost of individual computers easy to justify.

For complete information about this powerful breakthrough in 32-bit computing, contact the local HP sales office listed in your telephone directory. Ask a Technical Computer Specialist for a demonstration. Or write to Pete Hamilton, Dept. 41151, Hewlett-Packard, 3404 East Harmony Road, Fort Collins, CO 80525.



HEWLETT PACKARD

UNIX is a trademark of Bell Laboratories. Ethernet is a trademark of Xerox Corporation.

Speak German like a Diplomat!

What sort of people need to learn a foreign language as quickly and effectively as possible? *Foreign service personnel,* that's who.

Now you can learn to speak German with the same materials used by the U.S. State Department — the Foreign Service Institute's *Programmed Introduction to German.*

The FSI spent thousands of dollars and many years developing these materials for use by members of the United States diplomatic corps. Today people in all walks of life who need to learn to speak a foreign language are turning to this outstanding audio program.

The FSI's German Course is by far the most effective way to learn German at your own convenience and pace. It consists of a series of cassettes, recorded by native German speakers, and accompanying textbook. You simply follow the spoken and written instructions, listening and learning. By the end of the course you'll find yourself learning and speaking entirely in German!

This course turns your cassette player into a "teaching machine." With its unique "programmed" learning method, you set your own pace — testing yourself, correcting errors, reinforcing accurate responses.

This Programmed Course comes in two volumes, each shipped in a handsome library binder. You may order one or both:

☐ **Volume I. Programmed Introduction** to German, 10 cassettes (13 hr.), 647-p. text, $125.

☐ **Volume II. Basic Course Continued,** 8 cassettes (8 hr.), and 179-p. text, $110.

(Conn. and N.Y. residents add sales tax.)

TO ORDER BY PHONE, PLEASE CALL TOLL-FREE NUMBER: **1-800-243-1234.**

To order by mail, clip this ad and send with your name and address, and a check or money order — or charge to your credit card (AmEx, VISA, MasterCard, Diners) by enclosing card number, expiration date, and your signature.

The Foreign Service Institute's German course is unconditionally guaranteed. Try it for three weeks. If you're not convinced it's the fastest, easiest, most painless way to learn German, return it and we'll refund every penny you paid. Order Today!

81 courses in 36 other languages also available. Write us for free catalog. Our 10th year.

Audio-Forum
Suite 18 J
On-the-Green
Guilford, CT. 06437
(203) 453-9794



AUDIO-FORUM®

the underlying sea, giving rise to "an offshore pressure gradient which, in concert with offshore wind stress, reversed the mean flow."

The other group, consisting of D. R. Topham and R. G. Perkin of the Institute of Ocean Sciences in British Columbia, S. D. Smith and R. J. Anderson of the Bedford Institute of Oceanography in Nova Scotia and G. den Hartog of the Atmospheric Environment Service in Ontario, investigated the polynya in Baffin Bay between Devon Island and Dundas Island. The area is notable for strong tides and currents; in addition the narrowest part of Pioneer Channel, which runs between the islands, proves to be bridged "by a shallow sill at a depth of approximately 40 meters." The workers measured ocean currents, temperature and salinity in March and April, 1980. The temperature varied with the tides by as much as .2 degree Celsius, resulting in "pulses" of ocean water. The colder pulses were "close to the local freezing point." They "probably originate in shallow nearshore regions." The warmer pulses rose to the surface at the south side of the channel. They appear to have derived from "strong tidal flows across [the] shallow sill." Their slightly higher temperature is sufficient to establish the polynya.

Help!

As many a tale of adventure fiction has it, the best way for a hunter to attract a tiger to his blind is to tether a young goat nearby. The bleating kid, frightened by restraint and darkness, is irresistible tiger bait. Goats are by no means the only mammals that scream with fright; so do, among others, rabbits, lambs, horses and human beings. The fear scream, however, is particularly common among birds; in many species cries of fright by young birds quickly attract a parent.

What is the adaptive value of such screams? The question was framed and answered recently by Göran Högstedt of the University of Lund. Writing in *The American Naturalist,* Högstedt dismisses one interpretation: that the victim's shrieks serve to warn others of its kind that a predator is present, so that its behavior can be classed as a form of altruism. Observations show that most birds actually ignore the fear screams of their own kind. Is screaming juvenile behavior, as with a threatened nestling, that has somehow persisted in the adult? Probably not. Frogs and toads that grow up without interacting with their parents show the same fear-scream behavior. Does screaming startle the predator, giving the prey a chance to escape? This is a speculation along the right line, Högstedt notes, but one that does not go far enough. What the scream actually achieves is the attraction of other predator species to the site of an attack in the hope of an easy meal: the prey, the original predator or both.

To test his proposal Högstedt recorded fear screams of the European starling and played them in the field when potential second predators happened to approach within 100 to 250 meters of his concealed loudspeaker. The most numerous predators he saw were common buzzards. Of the 32 individuals of the species that came into range, 14 quickly approached the source of the recorded screams. So did one out of three rough-legged buzzards that came within range, two out of four sparrowhawks, four out of 10 hen harriers and a single goshawk. The only predatory birds that failed to respond to the screams at all were seven kestrels. Adult kestrels weigh only 200 grams, compared with a goshawk's 1,300 grams; they may have been avoiding approaching larger predatory birds. The starling screams also attracted two predatory mammalian species: all of the four foxes that came within range and two of the three feral cats.

The field test left no doubt that fear screams attracted diverse predators. Predatory birds, however, hunt more by eye than by ear. In open terrain they would probably spot a predator attacking prey well before the prey's screams could be heard. This being so, the adaptive advantage of the prey's summoning a second predator should be manifested to a greater extent among birds that habitually live out of sight in dense cover and to a lesser extent among birds that live primarily on the wing or on open ground.

Högstedt tested this prediction by determining the proportion of fear-screamers among six species of aerial and open-ground birds and among six species of dense-cover birds. Of the six species in the first category (swifts, sand martins, swallows, house martins, tawny pipits and white wagtails) two did not scream at all, and on the average only 5 percent of the individuals of the remaining four species proved to be fear-screamers. Of the six species in the second category (wrens, thrush nightingales, marsh warblers, reed warblers, icterine warblers and garden warblers) more than 80 percent of the individuals of two species were fear-screamers, and none of the other four species showed less than 32 percent fear-screamers.

Högstedt concludes that fear-screaming is not an altruistic adaptation but a selfish one, a "cry for help" directed at any animal capable of interfering with the primary predator. The value of the adaptation is the possibility—slight, but better than none—that a clash between a primary predator and a "pirate" one will facilitate the escape of the screaming prey.



How we lowered the cost of membrane keyboards so you can feel it.

Companies everywhere are looking for the technological leadership that will improve their quality and productivity.

Those looking to the future with electronic keyboards are finding it with AMP technology.

The way we design membrane keyboards to work with either flat or keybutton designs, or our use of the latest conductive ink technology to stand up under millions of actuations—these are examples of that leadership.

Add the recognized advantage of AMP interconnection systems, and our customers get the mass production and reliability needed to accommodate today's electronic technology. And at a cost that's lower than conventional systems.

How can we help you?

Whether you need keyboards, connectors or the technology for either, come to us and you'll get what the world has come to expect. Industry-leading products that are designed to reduce the cost of making yours.

AMP has a better way.

For product innovations, for new applications, for ways to save you money, AMP has the better ways to make your company more productive.

Write for our brochure, "Why AMP means productivity." AMP Incorporated, Harrisburg, PA 17105.

Auckland • Barcelona • Brussels • Buenos Aires • Dublin • Frankfurt • Harrisburg • Helsinki • s-Hertogenbosch • Kowloon • London • Luzern • Mexico City Montreal • Oslo • Paris • San Juan • Sao Paulo • Singapore • Stockholm • Sydney • Taipei • Turin • Toronto • Tokyo • Valley Forge • Viby • Vienna

AMP means productivity.

Learning in a Marine Snail

Hermissenda can be conditioned, or trained to associate two stimuli. Neural mechanisms underlying associative learning have been defined in the snail, and they may not be very different in the human brain

by Daniel L. Alkon

Can sophisticated mental functions in higher animals ever be accounted for in terms of the electrical activity of particular nerve cells and the chemical events that generate the activity? Neurobiologists can indeed describe some simple, stereotyped behaviors at that rigorously physical level. For example, the reflexive withdrawal of a limb in response to a harmful stimulus can be accounted for in terms of the electric impulses of a few identifiable cell types; many of the electrophysiological and biochemical mechanisms underlying the firing of those impulses are known. There is an immense gulf between a reflex action and, say, human memory and learning. Yet there is at least a chance that one can begin to bridge the gulf through the analysis of simple forms of learning in animals with a simple nervous system.

Pigeons, dogs and human beings can be conditioned, that is, they can be trained to associate one stimulus with another temporally related stimulus. The dog learns—and remembers—that the sound of a bell precedes and presages the smell of meat, and so it salivates at the sound of the bell. The vertebrate brain is far too complex to allow an analysis of such associative learning at the level of individual cells with present-day technology. For several decades, however, neurobiologists have been studying, in invertebrate animals, small networks of nerve cells that are responsible for integrating sensory information; in the past 20 years some have gone on to study the neural circuitry underlying simple reflexive behaviors. Some 12 years ago my colleagues and I at the Marine Biological Laboratory in Woods Hole, Mass., wondered whether such animals, with many orders of magnitude fewer neurons (nerve cells) than the trillion or so cells in the human brain, might be susceptible to conditioning. If we could train such an animal to associate two stimuli, we would try to identify the neural pathways activated by those stimuli and find out how the pathways converge, providing the physical basis for the learned association. By studying the activity of neurons in the pathways we would then try to identify specific changes effected by the conditioning and so begin to define "learning" in electrical and molecular terms.

Training

We chose to work with a gastropod mollusk, the shell-less marine snail *Hermissenda crassicornis*. In its natural environment *Hermissenda* exhibits positive phototaxis during the daylight portion of its light-dark cycle, that is, it moves toward light. The adaptive value of this behavior is clear: hydroids, the microorganisms on which the sea snail feeds, are concentrated in well-lighted water near the surface. When the ocean becomes turbulent, as in the course of a storm, *Hermissenda* responds by decreasing its overall velocity of movement and increasing the surface area of its muscular "foot." These responses enable it to cling to hard surfaces deeper in the water. Again the survival value is clear. An animal approaching the surface during a storm is shaken vigorously. It can survive for many weeks without feeding, but it cannot survive structural damage wrought by a turbulent sea.

Perhaps the experience of encountering turbulence in association with light at the surface causes the snail to "learn" to associate the light with the turbulence. The light would come to presage the onset of turbulence, and the learned association would tend to retard the animal's movement toward maximal light and thus into an area of maximal turbulence. We set out to reproduce the effects of light and turbulence in the laboratory.

We trained the snails in glass tubes filled with artificial seawater and arranged radially on a turntable. First we measured the average time untrained animals took to move to an illuminated area at the center of the turntable. Then we rotated the turntable at various speeds, subjecting the animals to a centrifugal force simulating the effects of turbulence. The conditions varied. For some animals light and rotation were precisely and repeatedly paired; five control groups were subjected to neither light nor rotation, to one or the other or to different combinations and temporal associations of the two stimuli.

After training we retimed the animals' movement toward light and calculated the effect of the training. For animals subjected to paired light and rotation (and only for those animals) the velocity of movement toward light was less than a third of the pretraining velocity. They had learned to associate light (which in this case served as the conditioned stimulus) with rotation (the unconditioned stimulus). The effect of the conditioning was to make the response to light more like the response to rotation.

Conditioning

Our various control procedures indicated that the change of light-elicited behavior after training truly reflected associative learning. Exposure to light alone did not change the subsequent response to light, ruling out a nonassociative behavioral change known as habituation: a progressive decrease in responsiveness resulting from the repeated presentation of a stimulus. Exposure to rotation alone did not change the subsequent response to rotation, ruling out the possibility of sensitization: an increase in responsiveness resulting from the repeated presentation of a strong motivational stimulus. (Both habituation and sensitization can confuse a finding that associative behavorial change has taken place; such effects have not been completely ruled out in the case of certain behavioral changes reported to follow associative training in other snail species.) Moreover, *Hermissenda* was not conditioned when light and rotation were alternated during training or when the stimuli were presented at random. Terry J. Crow and I found the two stimuli had to be presented together. (More specifically, the onset of light preceded



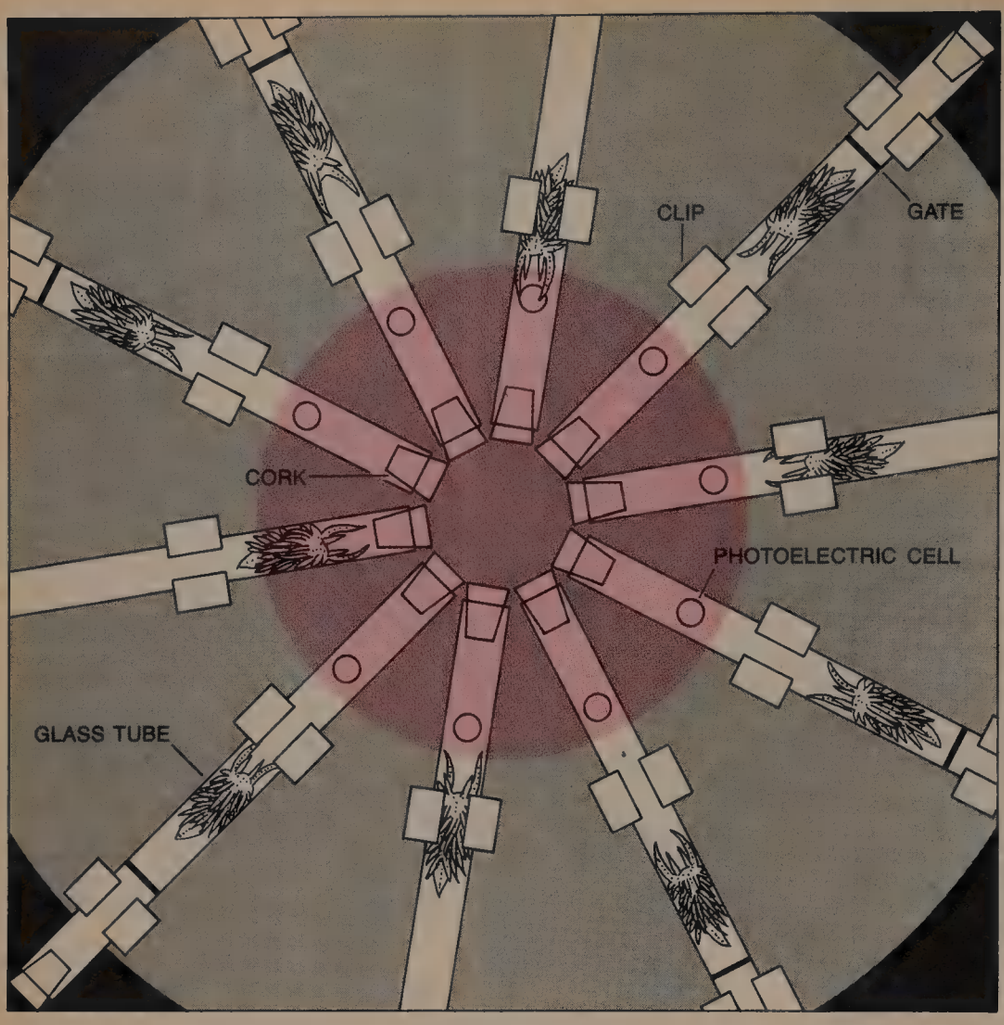
MARINE SNAIL *Hermissenda crassicornis,* **pictured in a seawater aquarium in this photograph made by Pierre A. Henkart of the National Cancer Institute, is about an inch and a half long. The two large oral tentacles are tactile and chemosensory organs; the smaller dorsal tentacles called rhinophores are thought to sense water movement. The animal's central nervous system is on the buccal crest just behind the rhinophores. The plumelike appendages covering the snail's back, the cerata, may serve, like gills, as gas-exchange organs.**



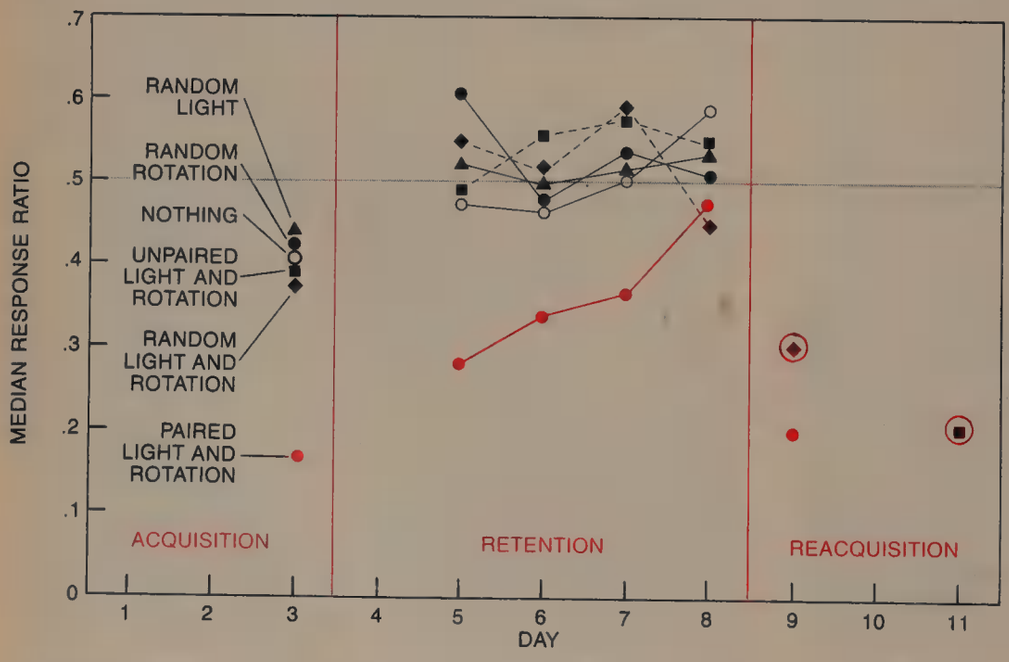
CENTRAL NERVOUS SYSTEM of *Hermissenda* **is enlarged some 50 diameters in a photomicrograph made by Alan M. Kuzirian of the author's laboratory. The methylene-blue stain picks out the globular ganglia, or clusters of related nerve cells, and a number of nerves radiating over the buccal crest. The eyes (*black*) and the primitive vestibular organs called statocysts (*white dots*) are between the two pedal ganglia (*lower left and right*) and the more medial pleural ganglia.**



EYE AND STATOCYST are enlarged about 300 diameters in a micrograph made by Kuzirian. The light-sensitive membranes of five photoreceptor cells extend up through the black pigment to abut on the underside of the lens. The roughly spherical statocyst wall is made up of 13 hair cells. Their motile cilia (not visible here) keep a cluster of crystals, called statoconia, centered; changes in the direction of gravitational force move the crystals, exciting particular hair cells.



SNAILS ARE TRAINED AND TESTED in glass tubes on a turntable. First each snail's phototactic response to light before training is measured: the time the snail takes to reach a spot of light (*color*) is recorded automatically when the animal reaches a photoelectric cell. Then the snails are trained by being rotated while confined to the outer end of the tube and thus subjected to a centrifugal force that is sensed by the statocysts. For one group the rotation is precisely paired with a 30-second period of light; various control groups either are not trained or are subjected to light or rotation alone or to the two stimuli alternately or at random. Finally the snails' velocity of movement toward light is timed again to determine the effect of training.



CONDITIONING IS DEMONSTRATED by these data. The "response ratio" measures the suppression of the phototactic response. A ratio of .5 means in effect that there has been no change in the velocity of movement toward light; a lower ratio signifies a reduction in the velocity. Animals subjected to paired light and rotation (*colored dots*) are conditioned after three days of training: they move more slowly toward light. They "forget" that behavior slowly but reacquire it after one day of retraining. Control animals show no significant change in behavior (*black*). Yet they too can be conditioned (*colored circles*) when subjected to paired stimuli.

the onset of maximum rotation by about a second.) It was the fixed temporal relation between stimuli that the animal learned and remembered.

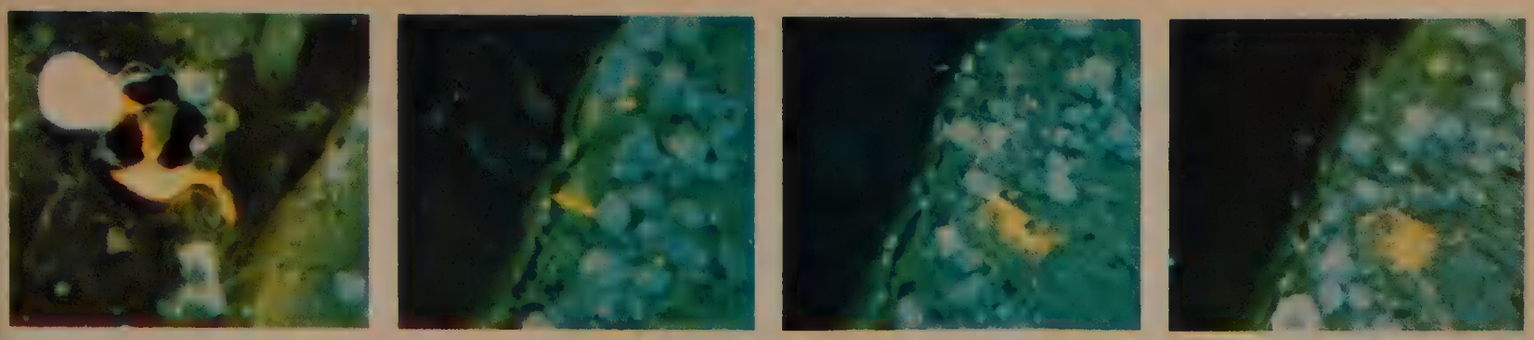
The behavioral change showed other characteristics of classical conditioning. It showed "acquisition" (the amount of change increased with practice), it could last for many weeks (close to a generation for this snail) and it did not affect the animal's response to stimuli other than the two being manipulated. William Richards and Joseph Farley of Princeton University showed that the learning was subject to extinction: after training was complete the animals "forgot" the learned association more quickly the more often light was presented without rotation. Crow showed, moreover, that after extinction the animals relearned the association with less training than they had the first time; the accelerated relearning is known as savings. Finally, Farley showed that *Hermissenda* learned more slowly when extra, unpaired stimuli (light or rotation) were interposed during conditioning. This degradation of learning by unassociated stimuli demonstrated contingency: the animals were learning that rotation is contingent on light, and when it was not, the learning process was impaired.

The behavioral changes I have described were observed in snails taken from their natural environment on the coast of California. Such animals displayed a wide range of responses to light before training, presumably reflecting individual differences in experience as well as in genetic background. June F. Harrigan therefore developed a laboratory strain of *Hermissenda* whose pretraining responses were much less variable, as were the behavioral changes observed in the course of training. We have also been able to raise animals with a defect in the organ that senses rotation and so obtain further evidence for associative learning. Such snails respond to light normally but cannot associate light with rotation.

All in all, these results suggested that *Hermissenda*'s ability to learn associatively could not be readily distinguished from the ability of a dog (or a whale or a human being, for that matter). This being the case, it seemed to us the underlying processes responsible for such learning might transcend differences among species. The first step toward understanding such processes was to trace the pathways mediating the association between light and rotation in the snail.

The Neural Pathways

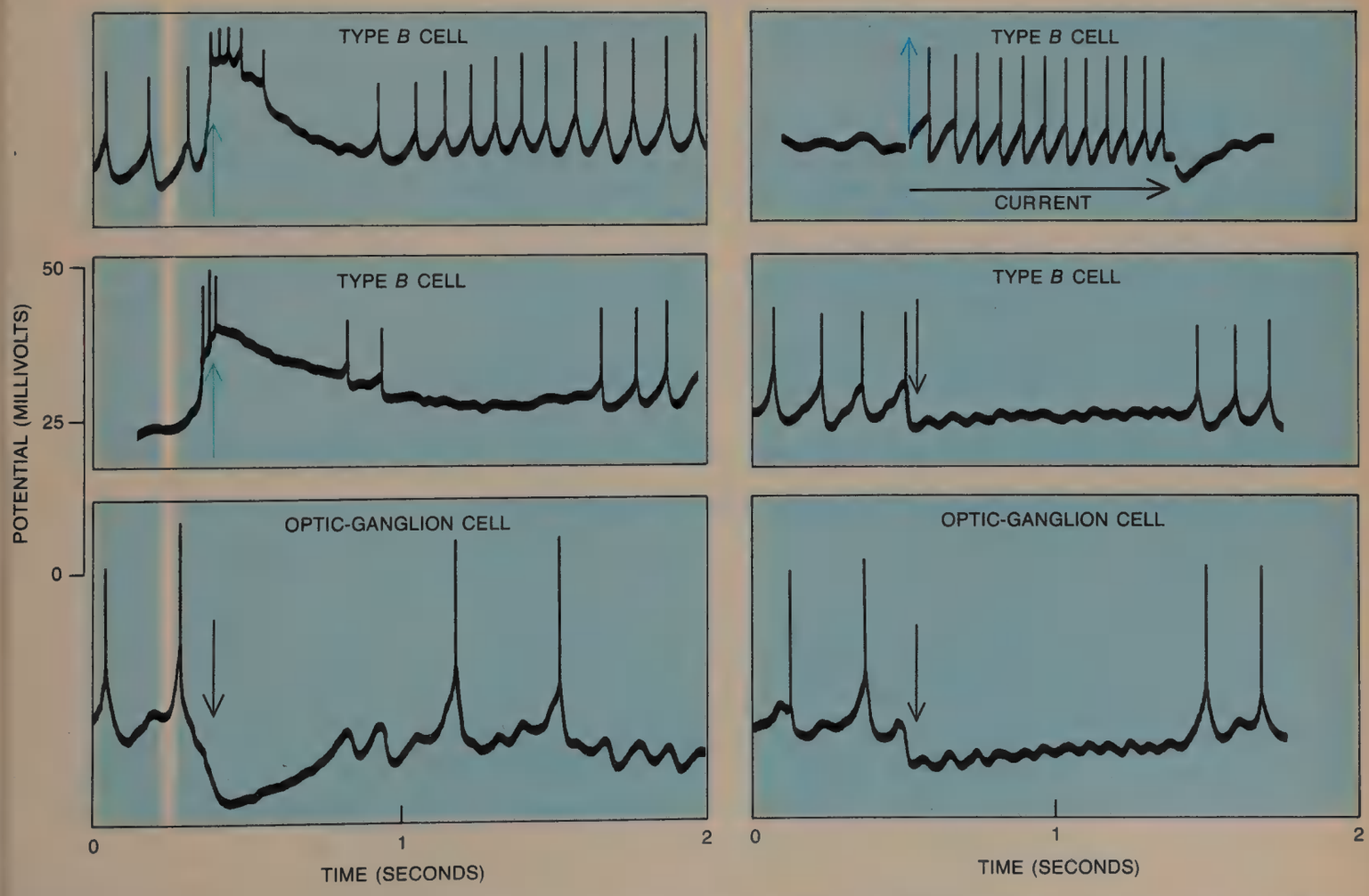
We began by mapping a wiring diagram of *Hermissenda*'s nervous system, with particular attention to two sensory pathways, one of them responsive to the conditioned stimulus (light) and the other to the unconditioned stimulus (rota-



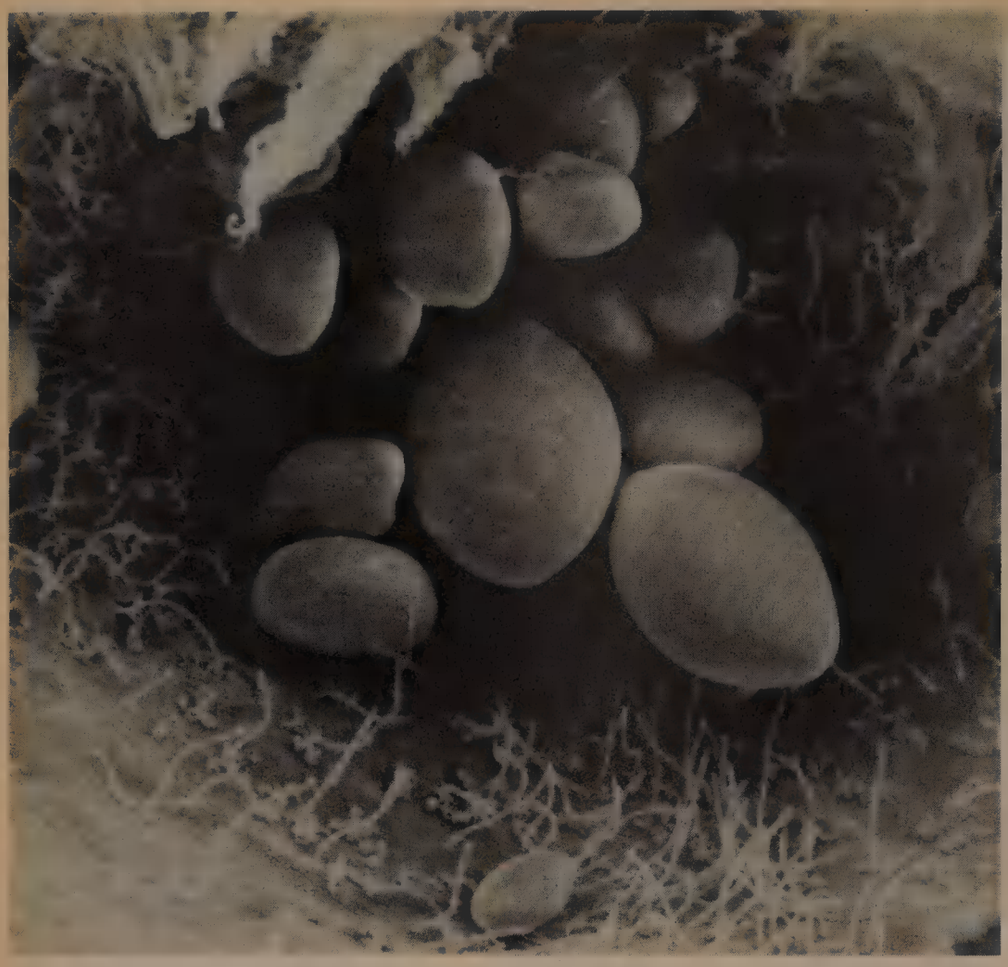
TYPE *B* PHOTORECEPTOR is mapped in micrographs made by Leona Masukawa. The dye Procion yellow was injected into a Type *B* cell and was transported along its axon to the axon terminals. In the first micrograph the dye fills the cell body (near the white lens) and the beginning of the axon. In successive thin sections of tissue the axon is seen (sectioned obliquely) entering the pleural ganglion and, in the last micrograph, ending in a spray of fine branches. The terminals of the branches form synapses (junctions) with other cells.

tion). Light generates electrical signals in two Type *A* photoreceptor cells and in three Type *B* photoreceptor cells in each of the animal's eyes; these signals are transmitted sequentially to interneurons, motor neurons and groups of muscles. Rotation is transduced into electrical signals by 13 hair cells (named for the hairlike cilia that line their inner surface) in each of two primitive vestibular organs analogous to the mammalian inner ear: the statocysts. These signals too are transmitted successively to interneurons, motor neurons and muscle groups. Some terminals of hair-cell axons (an axon is the long main process that carries the nerve impulse from one neuron to others) have synapses (junctions) with the axons of visual receptor cells. These and other, indirect interconnections between the visual and the vestibular pathways are the loci of convergence between the two sensory modalities.

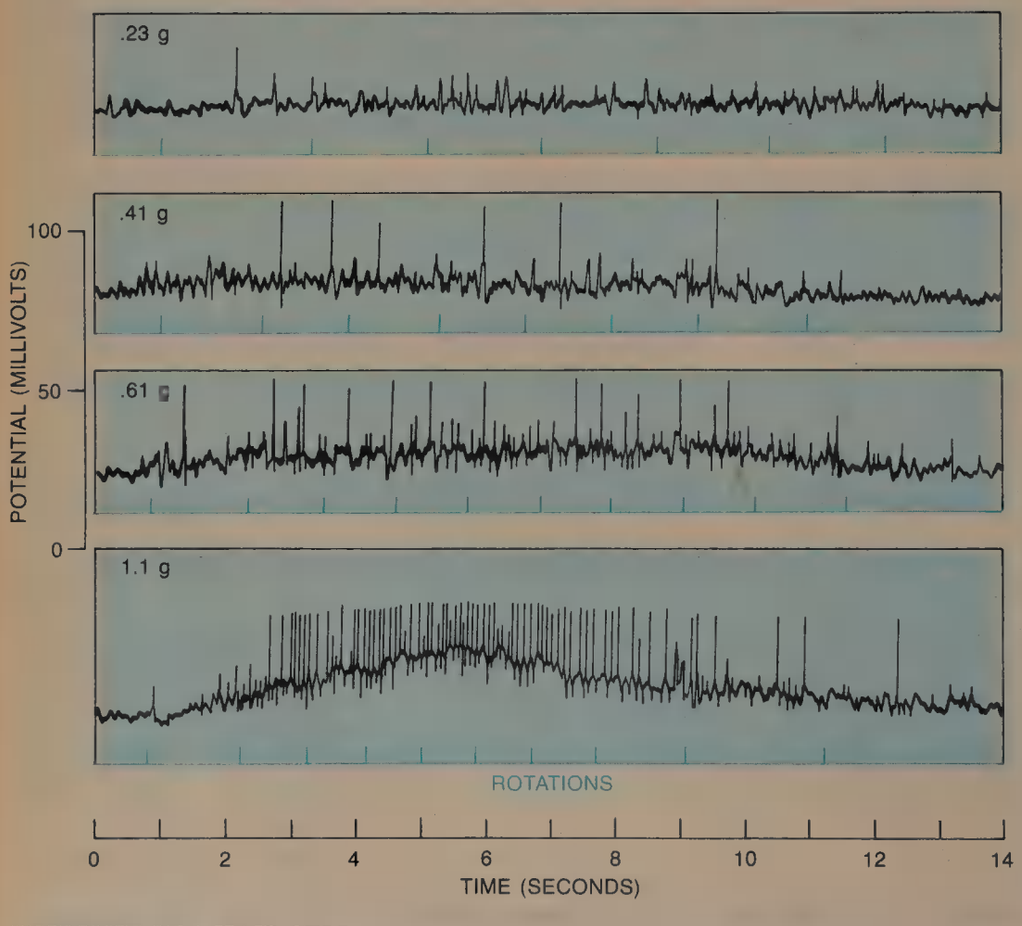
We mapped the two pathways both by electrophysiological methods and by making micrographs of stained preparations. The electrical responses of cells to light or rotation (or to the injection of microcurrents simulating the natural stimuli) were recorded with microelectrodes. By stimulating and recording from pairs of cells in thousands of animals we learned when and how specific receptor cells respond to particular stimuli by firing nerve impulses along the length of their axon; we learned to which postsynaptic neurons the signal is transmitted and whether it excites or inhibits such neurons. Repeating the process for successively connected cells enabled us to work out the pathways from the receptors that sense light and rotation to the muscles that impel *Hermissenda* through the water. The electrophysiological data were supplemented



INTRACELLULAR RECORDINGS show the effect of light (*left*) and of a current injection (*right*) on Type *B* photoreceptors and hence on an optic-ganglion cell they inhibit. In response to a brief flash of light (*white strip*) two *B* cells become depolarized, or more positively charged (*colored arrows*). Nerve impulses are triggered. These signals generate inhibitory postsynaptic potentials in the optic-ganglion cell, making it more negatively charged (*black arrow*); fewer impulses are triggered. Injection of a positive current (*horizontal arrow*) into either of two *B* cells simulates the effect of light, eliciting impulses that inhibit both the other *B* cell and the optic-ganglion cell.



HEMISPHERE OF STATOCYST is enlarged some 2,000 diameters in a scanning electron micrograph made by Eliahu Heldman. One sees the inner surface of the hair cells and a number of the statoconia (dispersed here because they are not kept clustered by the motion of the cilia).



RESPONSE OF A HAIR CELL was recorded as an isolated nervous system was being rotated at various speeds. Centrifugal force (indicated as a gravitational force, *g*) drove the statoconia against the cilia, making the hair cell's membrane potential more positive and giving rise to impulses. The cell's axon had been cut in order to eliminate any synaptic effects on the cell.

by information from photomicrographs and electron micrographs. A dye injected into a cell body moves along the axon to its farthest terminals, and the stained fiber can be traced in successive thin tissue sections; when two communicating neurons are stained, the synaptic junction between them can be visualized.

By these means we built up a wiring diagram: an interconnected network of chains of linked neurons [*see illustration on page 76*]. The diagram was essentially the same in every adult *Hermissenda* we studied. Clearly it is specified genetically; it is the invariant substrate for whatever changes are brought about by a learning process. How does its activity change with conditioning? Our investigation of the changes was made at three successively more localized and specific levels.

The first level was neuroanatomical. We measured the changes, during and after conditioning, in the signals that arise within individual neurons and are transmitted to other cells. Then we pinpointed the neurons in which the signals change first and those in which the signals remain changed, thereby identifying the events that actually cause the acquisition and the retention of the conditioned behavior. At a second, biophysical level we dealt with the cell membrane of the neurons in which the critical changes take place. We found differences in the currents passing across those membranes—in the flux of specific ions, or electrically charged atoms—that account for the changed signals. Finally, at the biochemical level, we identified chemical reactions that control the altered flow of ions.

The Excitable *B* Cells

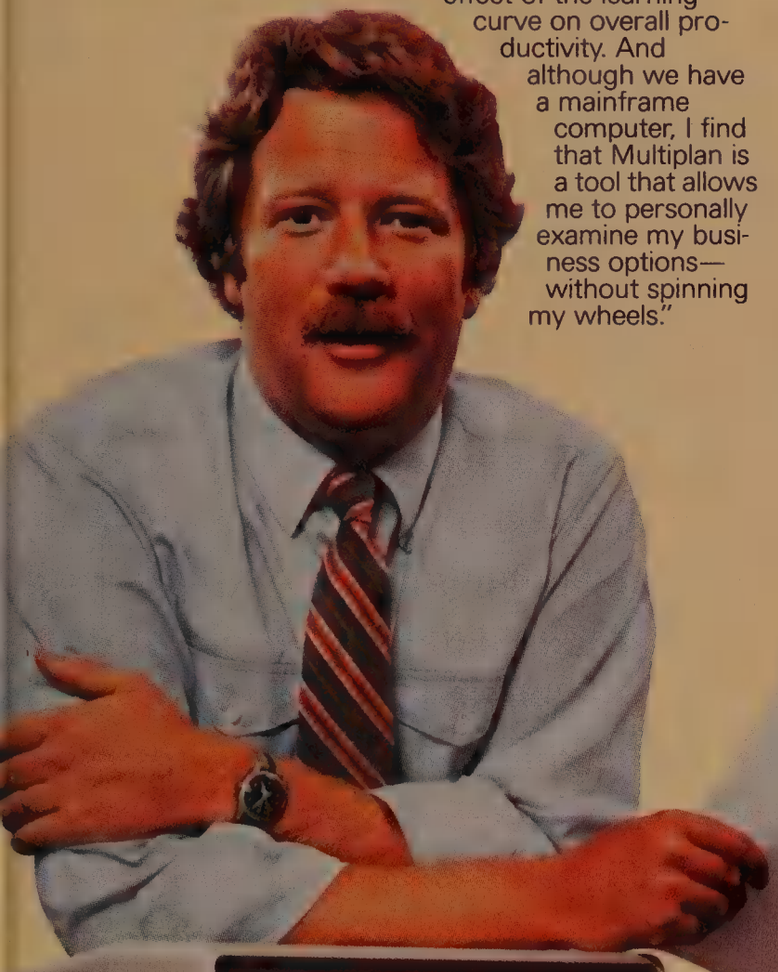
In essence our neuroanatomical study revealed a circuit of neurons whose activity is modified depending on the pattern of stimuli to which it is exposed. The potential for this redesigning is implicit in the wiring of the visual and vestibular pathways and their interconnections. When light is paired with rotation during training, the net response of the system is to excite the three *B* photoreceptor cells in each eye more than usual and to render them more easily excited. That is, the excitation of the Type *B* cells that normally results from stimulation by light is enhanced and prolonged after training.

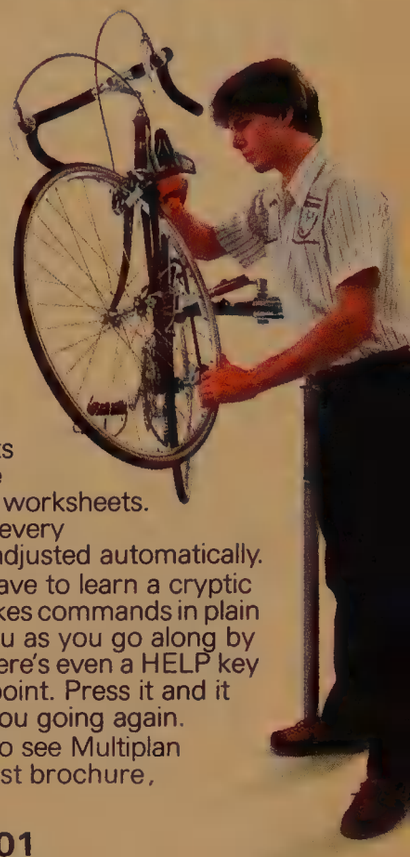
This is not a direct effect but a result of the integrated response of the system to the pairing of light and rotation. Normally the hair cells that respond most strongly to rotation fire even in the absence of excitation. This "tonic" activity sends inhibitory signals to the Type *B* cells. When the hair cells are stimulated by rotation alone, their activity is increased and they inhibit the *B* cells more strongly. When rotation stops, however,

"Microsoft® Multiplan™ allows me to explore more alternatives in less time. I think it leads to better management decisions."

Edward R. Schwinn, Jr.
President, Schwinn Bicycle Company

"Multiplan software helps me make better business decisions. It's as simple as that. With Multiplan on my microcomputer, I can explore a number of alternatives faster. That's really important to me in planning and tracking my totally hand-crafted Schwinn Paramount Bicycle. As with anything handmade, cost control and product planning are major considerations. For instance, Multiplan lets me study the effect of the learning curve on overall productivity. And although we have a mainframe computer, I find that Multiplan is a tool that allows me to personally examine my business options—without spinning my wheels."





A tool for decision makers. Microsoft Multiplan is a microcomputer software program that can help you too. In many ways. You can analyze cash flow. Plan budgets. Forecast income. Manage production. Multiplan allows you to set up an electronic worksheet for whatever your need may be. And, it lets you explore alternatives faster because when you make one change you immediately see the results of that change throughout the worksheet and on any related worksheets. Just change one number and every number that depends on it is adjusted automatically.

In plain English. You won't have to learn a cryptic language to use Multiplan. It takes commands in plain English. And, it will prompt you as you go along by telling you what to do next. There's even a HELP key in case you need help at any point. Press it and it gives you information to get you going again.

See it in action. If you'd like to see Multiplan in action, and receive our latest brochure, call us toll-free:

DEPARTMENT NO.01
1-800-547-3000
In Oregon, call collect 503-244-5500

We'll gladly locate the Microsoft dealers nearest you. They'll demonstrate Microsoft Multiplan and explain its compatibility with most popular personal computers. Including Apple® and the IBM® PC. It's compatible, in fact, with any computers that use the CP/M®-80 or MS™-DOS operating systems. Microsoft Multiplan is a good management tool for building bicycles. It's also a good management tool for building businesses. Like yours.

BETTER TOOLS FOR MICROCOMPUTERS

MICROSOFT™

Microsoft is a registered trademark, and Multiplan, MS, and the Microsoft logo are trademarks of Microsoft Corporation.



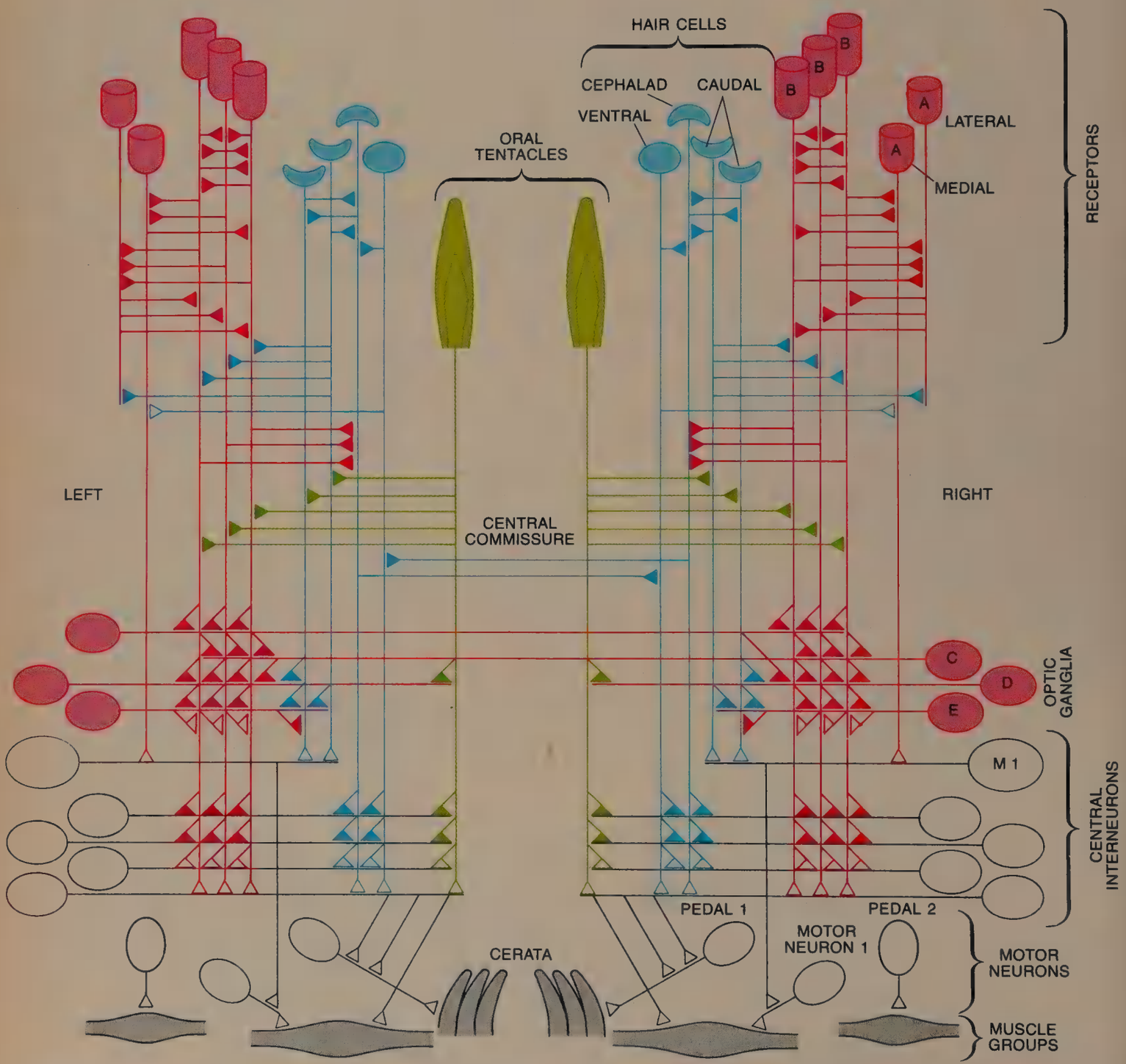
hair-cell activity is reduced even below the unstimulated, tonic level and the *B* cells are released from inhibition. When rotation is paired with light, there are additional effects. Currents that flow when the *B* cells transduce light reduce the electrical resistance of the photoreceptors' membrane, "shunting" the inhibitory signals from the hair cells and thereby lessening the effect of rotation-induced inhibition. Moreover, excitatory signals from second-order visual cells (the *E* cells of the optic ganglion) to the Type *B* cells increase in frequency immediately after rotation and light are paired. The effects of enhanced excitation accumulate in the course of training and the *B* cells become tonically more excitable: their response to light is progressively enhanced and prolonged.

Now, a salient effect of *B*-cell activity, as Yasumasa Goh showed in our laboratory, is to inhibit a chain of neurons that begins with the medial Type *A* cell and is responsible for motor-neuron impulses driving the muscle contractions that move the animal toward light. An enhanced and prolonged *B*-cell response has the effect of increasing the inhibition and thereby reducing the movement toward light.

Causality

A number of experimental findings demonstrated that the modification of the Type *B* cell's response is not merely correlated with the learning behavior but is at least in part responsible for causing the behavior. These confirming experiments were "blind," that is, the investigators did not know which training procedure particular animals had undergone. In one study we measured the activity of motor neurons mediating the turn toward light. As predicted, the motor neurons of conditioned animals (in whom the inhibitory input originating with the Type *B* cells would be expected to be enhanced) indeed showed reduced activity when the animals were exposed to light.

In another experiment on conditioned and control-group animals during the



WIRING DIAGRAM of *Hermissenda's* **nervous system is summarized schematically. Representative cells (and their axons) of the visual (*red*) and vestibular (*blue*) pathways are shown, as are some axons of the tactile-chemosensory pathways (*green*) and some interneurons, motor neurons and muscles (*black*). Excitatory synaptic interactions are indicated by open triangles, inhibitory ones by solid triangles. Each interaction has been shown to be reliably present by repeated simultaneous recordings of presynaptic and postsynaptic potentials.**



Even when it's empty, it's loaded.

The 1983 Honda Civic 4-Door Wagon is full of advantages. Look at its cargo area. It's fully carpeted. Easily accessible through a wide-opening liftgate. And offers maximum flat area loadspace with the rear seat folded down.

Like all Civics, the Wagon is also packed with quality Honda engineering features. Like our proven front-wheel drive and power-assisted, self-adjusting, ventilated front disc brakes.

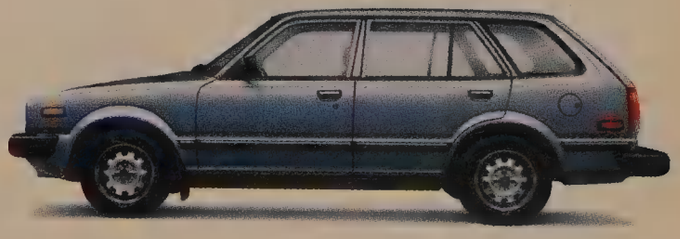
And thoughtful new additions. Like child-proof rear door locks, body side mouldings, a remote liftgate release and rear window wiper and washer.

The reasonable price and its overall economy can help lighten your load, too.

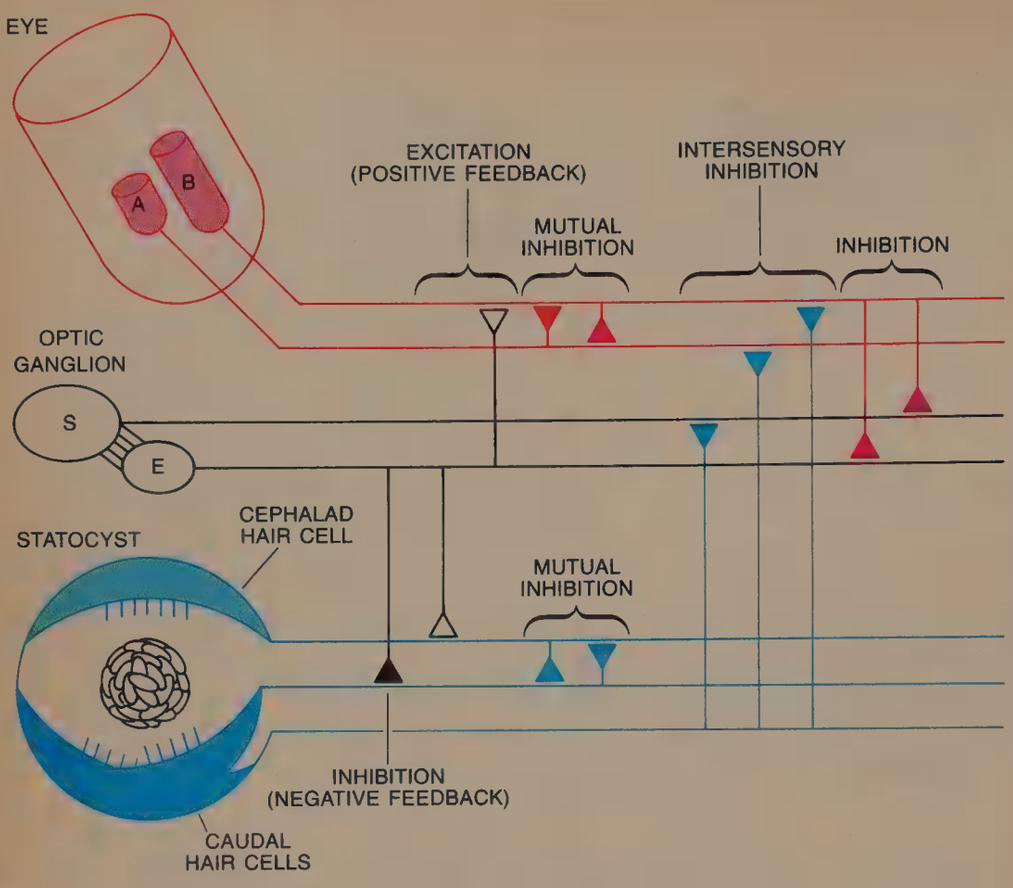
To appreciate our Wagon fully, come in and see it empty.

HONDA

We make it simple.



© 1983 American Honda Motor Co., Inc.



SITES OF CONVERGENCE between the visual and the vestibular pathways are indicated in this diagram of representative cells responsive to light and rotation. They include a "silent" optic-ganglion cell (*S*) whose effects are transmitted by way of a cell (*E*) to which it is coupled electrically. The caudal hair cells exhibit steady ("tonic") activity tending to inhibit Type *B* photoreceptors. Paired presentation of light and rotation is followed by reduced hair-cell inhibition of the Type *B* cell and by increased synaptic excitation of it by the *E* cell (positive feedback).

period of retention, the axon of Type *B* cells was cut close to the cell body, which was thus isolated physically and electrically from the rest of the nervous system. The excitability of the isolated cells in response to light or to the injection of a positive current was assessed. Several different indexes of excitability were increased in the Type *B* cells of the animals that had been exposed to paired light and rotation. The conditioning-induced increase in *B*-cell excitability, in other words, is intrinsic to (and is stored in) the cell body; it is not a passive reflection of changes elsewhere in the system.

Farley and I also demonstrated causality by manipulating the snail's neural circuits artificially. By injecting microcurrents into individual cells of the two converging pathways we caused the cells to send the same signals they send in response to the natural stimuli. In one experiment we worked with the nervous system isolated from the snail's body, injecting currents into a hair cell and a second-order visual cell of the optic ganglion. Repeated current injections, paired with light, gave rise to increased *B*-cell excitability; unpaired with light, the same injections (which now simulated one of our control procedures in training) had no such effect. In another experiment we achieved the behavioral effect of conditioning simply by pairing light with repeated injections, into a single Type *B* cell exposed in living animals, of a current calculated to simulate the synaptic effects on a *B* cell when hair cells are stimulated by rotation. We found that *B*-cell excitability was duly increased. On days following the injection, after the animals had recovered from the procedure, the velocity of their movement toward a light source was reduced, just as it was in the case of conditioned animals. Control animals that received the current injections unpaired with light displayed no such change.

A final demonstration of the causal role of the increased excitability of Type *B* cells came from confirmation of a behavioral prediction. The connections among visual and vestibular cells are such that rotating the animal with its head toward the center of rotation stimulates what are called the caudal hair cells, with the result that the Type *B* cells are eventually made more excitable, as described above. Training the animal with its head in the opposite direction instead stimulates what are called the cephalad hair cells and so results in increased inhibition of the Type *B* cells; the phototactic response, we reasoned, should be enhanced. We tried it and the prediction was borne out: the Type *B* cells became less excitable and the animals moved more quickly toward light. (Shaking without restriction in orientation, as in a turbulent sea, has the net effect of exciting the Type *B* cells because the caudal hair cells have more direct and more effective synaptic connections with the photoreceptors than the cephalad cells do.)

Membrane Changes

Our neuroanatomic results constituted the first evidence for any animal, vertebrate or invertebrate, that lasting associative learning is a direct result of a change in the excitability of particular, identified neurons. Excitability is a property of the nerve-cell membrane, and so the obvious question became: What happens to the Type *B* cell's membrane in the course of conditioning? Before discussing our findings I need to explain briefly how changes in the neuronal membrane generate electrical signals. At rest a neuron—the cell body and its processes—is polarized: the inside of the cell is about 70 millivolts negative with respect to the outside. The potential difference is maintained by the concentration of ions within and outside the cell; the concentration is regulated by the selective permeability of the cell's membrane, which is studded with gated protein channels that are specific for each of several different ions. The major factors in establishing the membrane potential are positive potassium and sodium ions. There is a much higher concentration of potassium inside the cell and a much higher concentration of sodium outside. The membrane is normally much more permeable to potassium than it is to sodium, and so the potassium ions tend to leak out, leaving the interior of the resting cell negative.

Ionic channels in the cell membrane can open in response to a sensory stimulus (in the case of a receptor cell), to the effect of a chemical transmitter on the postsynaptic membrane or to a shift of membrane potential. When sodium channels open, sodium ions (and in the Type *B* cell also positive calcium ions) flow in, depolarizing the cell (making the inside more positive). Almost immediately this depolarization causes potassium channels to open, and the outflow of potassium acts to restore a negative resting potential. These changes of membrane potential can spread locally along the membrane and in some cases give rise to a much larger depolarization that can propagate regeneratively along the entire length of the axon; this is the action potential, or nerve impulse.

We found that the outflow of potassium ions (through what are called *A* channels and probably through other potassium channels too) during a depolarization induced by a current injection in the absence of light is reduced on days after conditioning; the inflow of calcium ions (through other channels) follow-



Reliable.

You can count on 3M diskettes. Day after day.

Just like the sun, you can rely on 3M diskettes every day. At 3M, reliability is built into every diskette. We've been in the computer media business for over 30 years. And we've never settled in. We're constantly improving and perfecting our product line, from computer tape and data cartridges to floppy disks.

3M diskettes are made at 3M. That way, we have complete control over the entire manufacturing process. And you can have complete confidence in the reliability of every 3M diskette you buy.

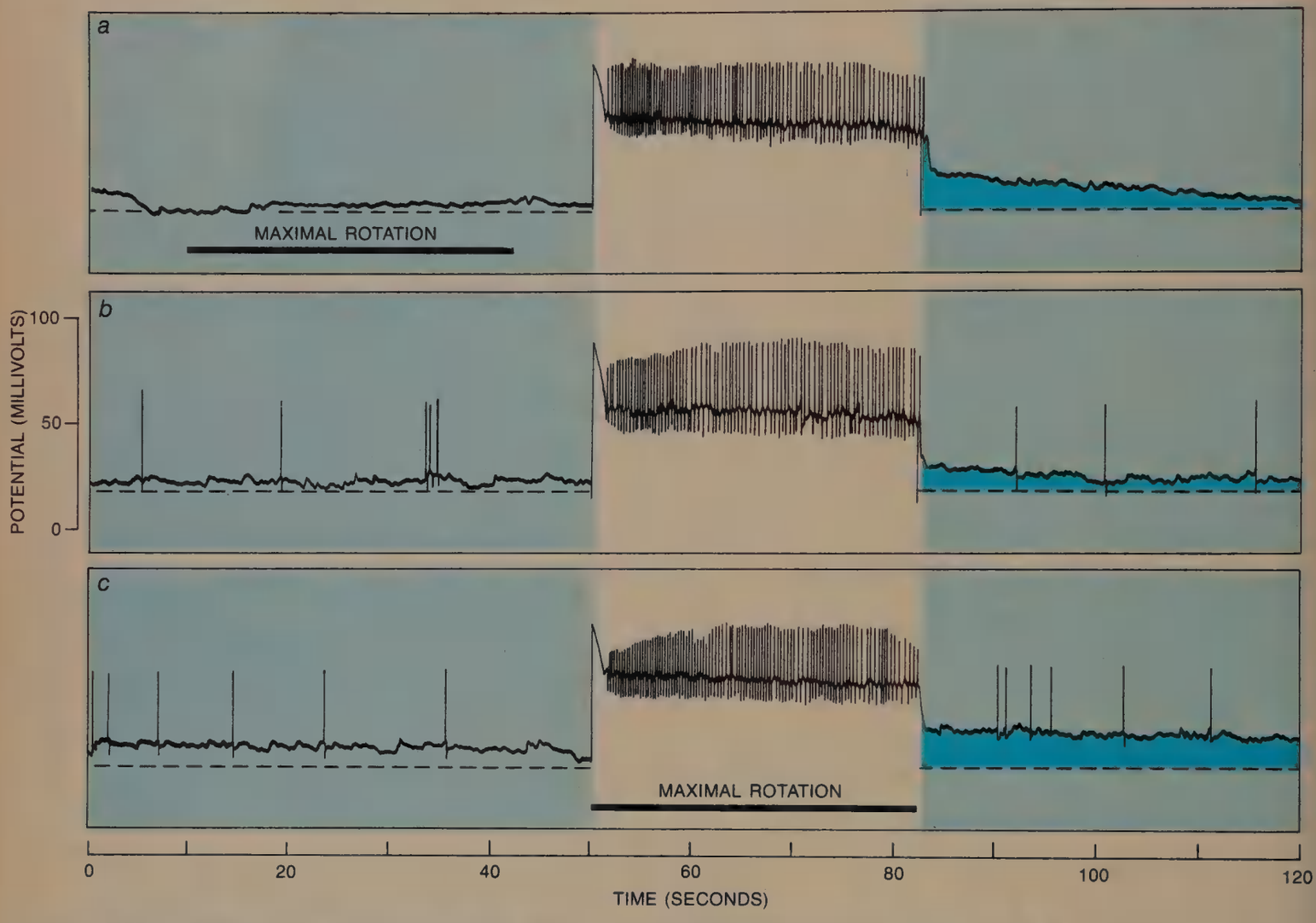
Look in the Yellow Pages under Computer Supplies and Parts for the 3M distributor nearest you. In Canada, write 3M Canada, Inc., London, Ontario. If it's worth remembering, it's worth 3M diskettes.



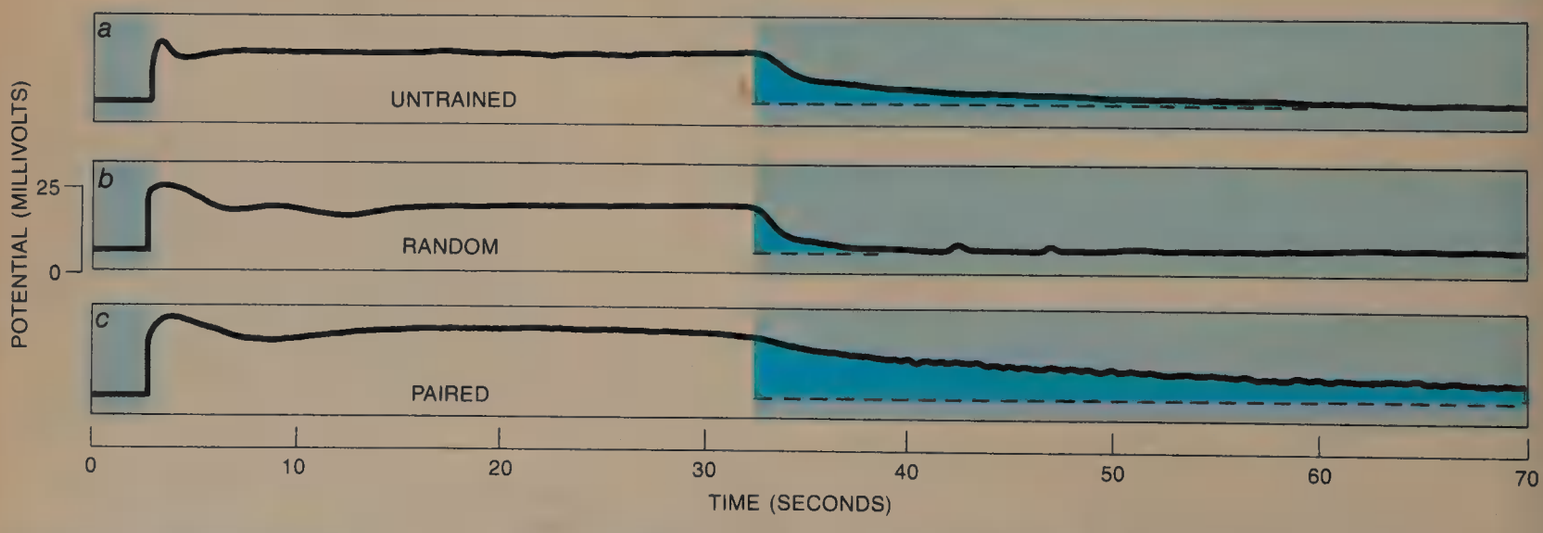
3M hears you...

3M

ing light-induced excitation is increased. Both of these changes cause the Type *B* cell to depolarize more in response to excitation by light or a current: it is more excitable. (It may be, although it is not yet established, that the increased calcium flux is simply the result of the increased excitability arising from the reduced flow of potassium.) John A. Connor of Bell Laboratories and I found, moreover, that calcium ions tend to accumulate within the Type *B* cell during the acquisition of the conditioned behavior. Prolonged elevation of intracellular calcium reduces the number of open *A* potassium channels. The reduced outflow of potassium makes the membrane more excitable. In the presence of light it depolarizes more readily; more calcium channels open, enhancing the flow of calcium and thereby further



TYPE *B* CELL'S RESPONSE to light alternating with rotation (*a*), light alone (*b*) or light paired with rotation (*c*) is shown. In each case the second of two successive periods of light is shown (*white area*); the first had ended 40 seconds before these traces begin. The cell's initial resting voltage is indicated by the broken line; the amplitude of the long-lasting depolarization that follows the second exposure to light is emphasized by the dark color. The paired stimuli can be seen to cause the most long-lasting depolarization of the Type *B* cells (*c*).



CHANGE IN EXCITABILITY is shown by these recordings to be a property of the Type *B* cell itself. The voltage of three *B* cells (whose axons had been cut to isolate them from all other neurons) was measured. The cells came from an untrained animal (*a*) and, on days after training, from an animal that had been subjected to light and rotation at random (*b*) and from one that had been conditioned by paired light and rotation (*c*). The long-lasting depolarization (*dark color*) following exposure to light is largest in the conditioned animal (*c*).

SCIENCE/SCOPE

Vacuum-tube computers spanning half an acre of floor space will be replaced by modern computers the size of two vending machines when North America's new air defense system goes into operation late this year. Hughes Aircraft Company's Joint Surveillance System will replace aging SAGE (Semi-Automatic Ground Environment) and BUIC (Back-Up Interceptor Control) systems. It will link U.S. Air Force surveillance radars, civil air traffic control radars, and Canadian radars into a shared system. Seven regional control centers -- each equipped with the smaller computers -- will monitor skies 200 miles beyond North American borders. An eighth center will monitor skies surrounding Hawaii.

The U.S. Army will save over $200 million by using simulators to train troops to use and repair Firefinder weapon-locating radars. The Firefinder detects and tracks enemy artillery and mortar fire with a pencil-thin electronic beam. It instantly backplots their trajectories so counterfire can be directed with pinpoint accuracy. The Firefinder trainer simulates battlefield conditions so troops can learn to operate the radar without using live artillery fire and without taking a radar out of deployment for instruction. Also, where only one student could operate an actual radar, six students can train at once under the control of one instructor. Hughes builds the Firefinder radars and trainers.

Mexico will inaugurate a national communications satellite system in 1985 with the launch of two Hughes spacecraft from NASA's space shuttle. The satellites will carry advanced telecommunications services for the entire country. Plans include educational and commercial TV programs, telephone and facsimile services, and data and business transmissions. Mexico now leases communications capacity on two other Hughes spacecraft -- an Intelsat IV from the International Telecommunications Satellite Organization and a Westar from Western Union. There are 157 satellite receiving stations operating throughout Mexico.

Paperless planning is making its debut to guide assemblers with step-by-step instructions for manufacturing electro-optical hardware. Color video monitors having twice the quality of home TV sets are replacing thick planning books at Hughes. The monitors are used in conjunction with video discs that hold up to 50,000 full-size color pictures, each equivalent to one sheet of planning. The discs store three-dimensional computer graphics or standard video. Assemblers can review still images or sequences showing how a product is to be built. Paperless planning is faster, more accurate, and less costly than conventional methods. It will also reduce manufacturing errors.

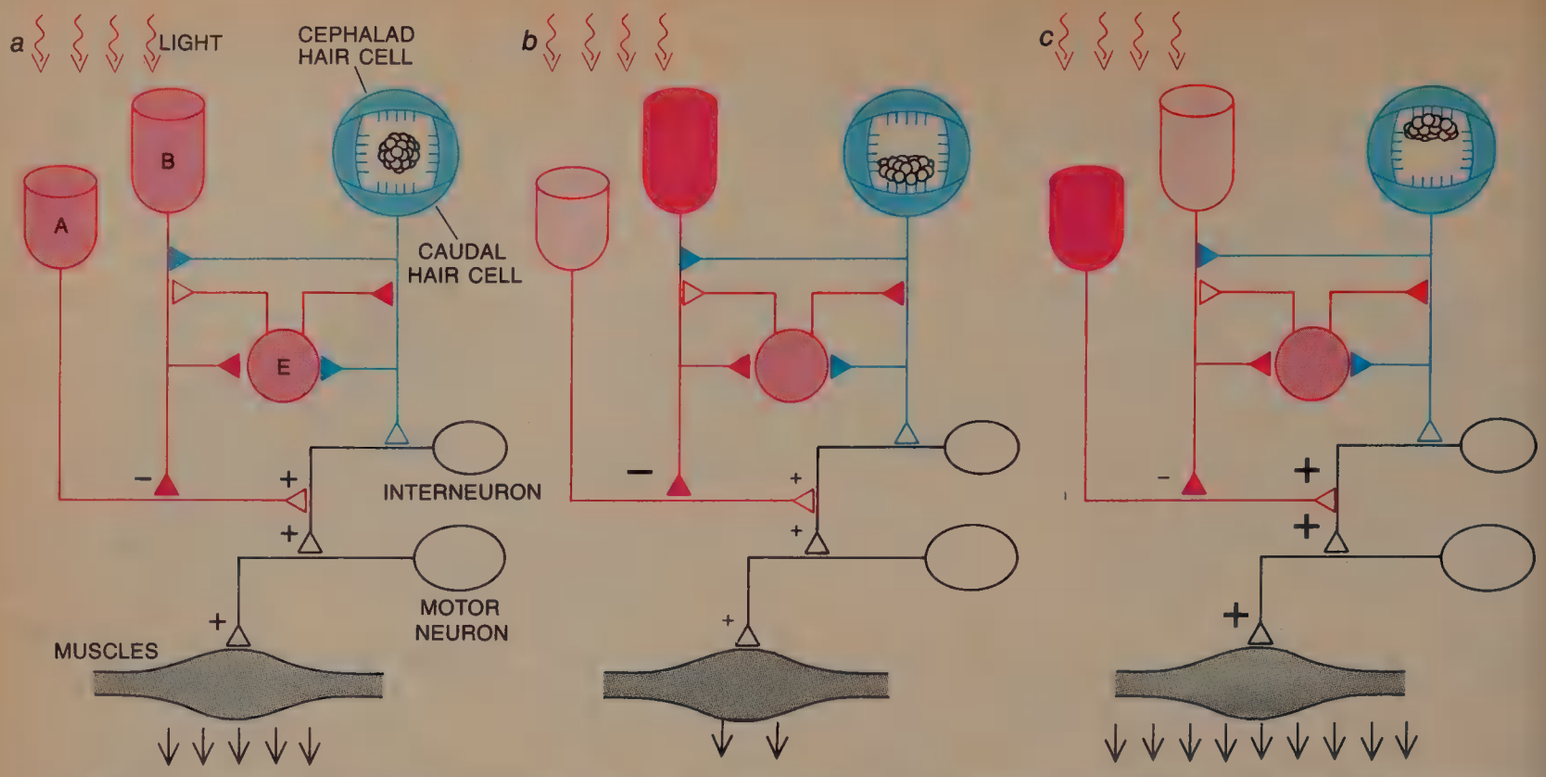
Hughes Missile Systems Group in Tucson, Arizona, has built a reputation for the development and manufacture of sophisticated electronics. We have enhanced our leadership in this field with these missiles: the electro-optical/laser/infrared-guided Maverick, the radar-guided Phoenix, the wire-guided TOW, and now the radar-guided AMRAAM. We have openings for microwave specialists, circuit designers, hybrid microelectronic engineers, and production support test engineers. Send your resume to Professional Employment, Dept. SE, Hughes Missile Systems Group, P.O. Box 11337, Tucson, AZ 85734. Equal opportunity employer.

Creating a new world with electronics

HUGHES

HUGHES AIRCRAFT COMPANY

For more information please write:
P.O. Box 11803, Los Angeles, CA 90291



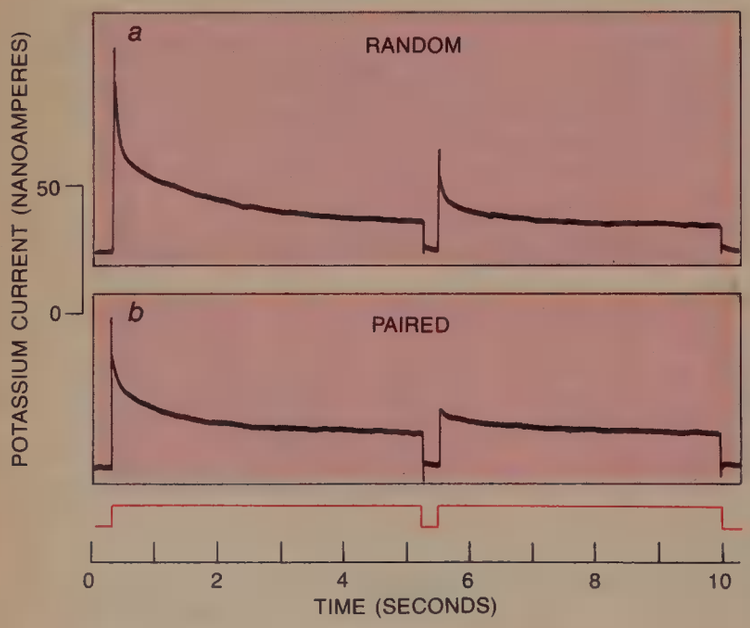
ANIMAL'S RESPONSE TO LIGHT is changed by the effects of conditioning on the neural system. The basic wiring diagram is invariant: the Type *B* photoreceptor inhibits the medial Type *A* photoreceptor, whose stimulation by light causes excitation of the interneuron, motor neuron and muscles in turn. During a burst of hair-cell impulses the *B* cell is inhibited (directly and by way of the second-order *E* cell). Following a burst the hair cell becomes quiescent and the *E* cell more active; the *B* cell is consequently excited. This is the situation in the untrained animal (*a*); the synaptic effects are enhanced when light and rotation are paired even once. Repeated pairing of light with rotation, with the snail oriented to excite caudal hair cells (negative conditioning), results in a long-lasting increase in the excitability of the *B* cell, increasing its inhibition of the Type *A* cell during and after exposure to light and thus slowing movement toward light (*b*). Positive conditioning (with the animal in the opposite orientation) excites cephalad hair cells and has the opposite effect (*c*).

increasing intracellular calcium, which further reduces the number of open potassium channels, and so on in a positive-feedback cycle.

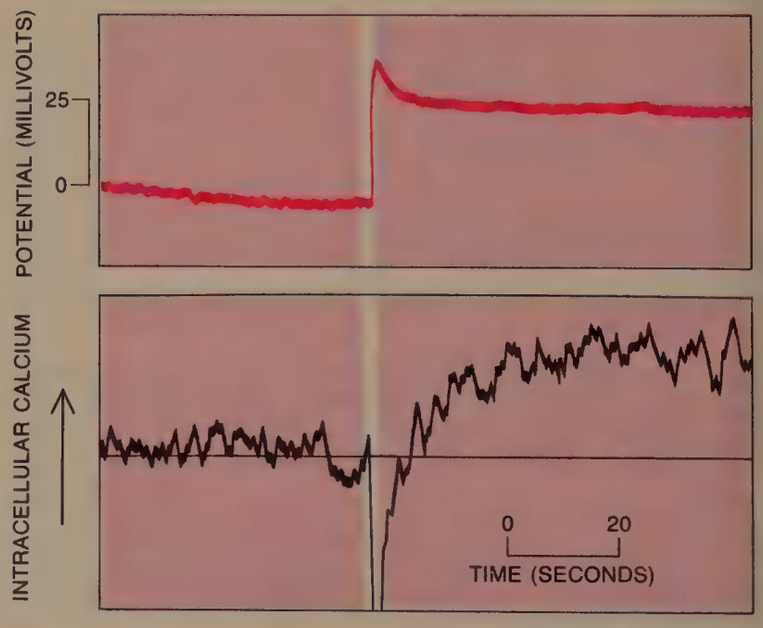
The cycle is reinforced during learning. With successive repetitions of stimulus pairings at sufficiently brief intervals, there is a progressively larger residual increase of Type *B* depolarization, which in turn seems to amplify the interaction of the calcium and potassium channels. It is possible too that the channels undergo complementary changes in synaptically linked neurons. We have preliminary evidence that the membrane of the medial Type *A* photoreceptor, the one inhibited by Type *B* cells, is less excitable after conditioning—the inverse of the situation in the Type *B* cells.

Biochemical Mechanisms

If the calcium level in Type *B* cells rises during conditioning and the rise



POTASSIUM OUTFLOW is reduced in the Type *B* cells of a conditioned animal on days after training. A change of potential (a 50-millivolt increase) was imposed on a *B* cell (*colored trace*) and the resulting current equivalent to the outflow of potassium ions was measured (*black traces*). The flow is larger in a cell from an animal exposed to light and rotation at random (*a*) than it is in a cell from an animal conditioned by paired light and rotation (*b*); the difference is greater for the second of the two successive changes in voltage.



INTRACELLULAR CALCIUM LEVEL is increased following a light stimulus. A Type *B* cell was injected with a dye that combines with calcium to form a complex absorbing light at a specific wavelength; the absorbance is recorded to serve as a measure of the calcium concentration in the cell. In response to a flash of light (*white strip*) the interior of the cell becomes more positive (*colored trace*). The calcium level (*black trace*) increases, and it remains elevated as long as the photoreceptor's membrane potential remains elevated.

is accompanied by changes in the flow of ions through membrane channels, we hypothesized, perhaps the calcium affects the channels by initiating or stimulating certain biochemical reactions. Calcium is implicated in many regulatory mechanisms, notably the activation of enzymes responsible for attaching phosphate groups to proteins; the phosphorylation in turn often changes the character of the protein. It seemed possible that elevated intracellular calcium activates enzymes that attach phosphate groups to membrane proteins controlling the opening and closing of *A* potassium channels and calcium channels.

A number of observations support the hypothesis. First Joseph T. Neary measured the incorporation of phosphate groups (labeled with a radioactive isotope) into certain proteins that appear to constitute or to be associated with potassium and calcium channels in the membrane. The net number of phosphate groups incorporated in one of those proteins is different in conditioned animals from what it is in control animals. Paul Greengard of the Yale University School of Medicine had found earlier that the phosphorylation of certain membrane proteins in several vertebrate species is mediated by calcium. Recently Leonard K. Kaczmareck and Susan A. DeRiemer of Yale, collaborating with Greengard and us, were able to link Neary's finding with Greengard's. They showed that in *Hermissenda* the phosphorylation of several nervous-system proteins, including the one Neary had found to change with conditioning, depends on the calcium concentration.

We tested calcium's primary role in increasing excitability by injecting the ion into Type *B* cells. The injection caused a prolonged reduction of potassium flow through the *A* channel, just as conditioning does. Finally, following a suggestion made by Howard Rasmussen of Yale, we injected into *B* cells a calcium-dependent protein kinase, an enzyme that catalyzes calcium-mediated phosphorylation. The injection enhanced the effect of calcium on the *A* channels. It simulated the effect of conditioning, giving rise to the same increased excitability due to a reduced flow of potassium ions and an increased flow of calcium ions.

In summary, our biochemical studies suggest that during acquisition the intracellular calcium level is progressively elevated. The ion activates calcium-dependent enzymes that facilitate the phosphorylation of membrane proteins regulating the flow of calcium and potassium ions. After training ends, intracellular calcium subsides to a normal level, but throughout the period of retention of learned behavior the phosphorylation of proteins persists. (As Rasmussen has pointed out, calcium-mediated phosphorylation can be a fair-

BIG BOOM FOR REAL ESTATE INVESTORS.

Thousands of people like you find new opportunities for income, appreciation and tax advantages through the CENTURY 21® network.

Tax laws have changed. And now real estate is an even better investment than ever before.

At CENTURY 21 offices all over the country, we're showing people how to take advantage of the latest tax laws. How they can use leverage. How they can now shelter more of their income. Even income from salary, stock dividends and other sources.

A CENTURY 21 Investment Professional can show you, too. Through real estate, you may be able to get a greater annual cash flow, after taxes, than through other traditional types of investment. And you don't have to invest a lot of time and effort. Because a CENTURY 21 office can also handle the details of property management for you.

The following chart illustrates the potential advantages of real estate over other investments:

	REAL ESTATE	MONEY MARKETS	STOCKS
Cash Flow	✔	✔	✔
Depreciation	✔		
Loan Amortization	✔		
Appreciation	✔		✔

The opportunities are impressive. And that's one of the reasons investment business is booming at CENTURY 21 offices. Last year, we helped investors buy or sell real estate worth over $4 billion. We're North America's leading real estate sales organization. And now, we're leading the way in real estate investment, too.

Century 21®

Call this toll-free number for a CENTURY 21 Investment Professional.

1-800-228-3399

Nebraska Residents please call 1-800-642-8788. In Alaska and Hawaii call 1-800-862-1100.

©1983 Century 21 Real Estate Corporation, as trustee for the NAF ® and TM-trademarks of Century 21 Real Estate Corporation Printed in U.S.A. Equal Housing Opportunity

EACH OFFICE IS INDEPENDENTLY OWNED AND OPERATED.

ly irreversible process.) The high level of intracellular calcium during training could cause a long-lasting activation, through phosphorylation, of protein kinases. The activated enzymes could then be responsible for the persistent membrane changes that maintain increased excitability.

Our findings demonstrate, for the first time for any nervous system, that associative learning lasting for several days involves changes in the long-term biochemical regulation of ion flow across the cell membrane. The changes we find are at cell bodies and along axons, not at synapses. It may turn out that synaptic changes are indeed implicated in associative learning, but we have not found evidence for such changes. Alterations of nonsynaptic membrane permeability can account for the learned behavior we have studied in *Hermissenda*.

Learning in Vertebrates

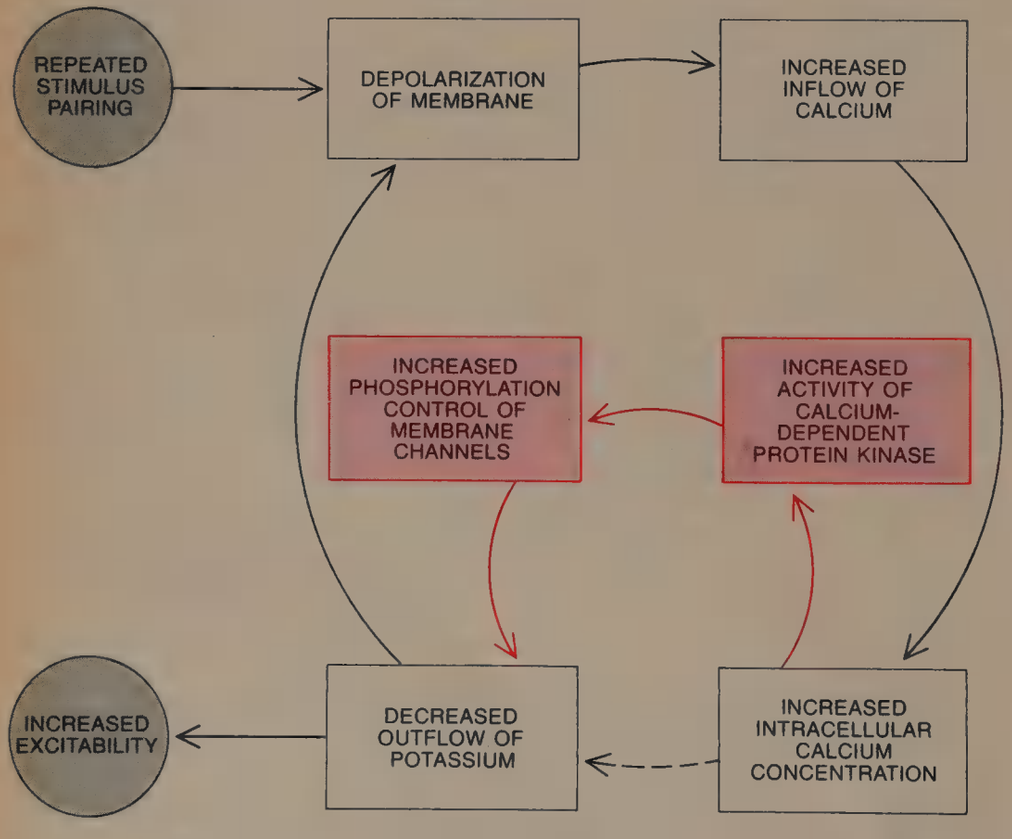
Marine organisms serve as marvelous models, revealing basic facts about many aspects of nervous function that have consistently proved to be generally applicable in higher animals. We have begun to test the applicability to the vertebrate brain of what we have learned about the biochemical basis of associative learning in *Hermissenda*. Recently Charles D. Woody of the School of Medicine of the University of California at Los Angeles, Bruce A. Haye and I injected the protein kinase I mentioned above (the enzyme whose injection into Type *B* cells simulated the effects of conditioning) into certain neurons in the brain of the cat. The activity of these neurons has been correlated with learning. The enzyme injection gave rise to changes in the excitability of the cat-neuron membrane that are similar to those generated in the Type *B* cell.

Such persistence through evolution of a common biochemical mechanism does not, to be sure, imply identity of function. Even if one considers only associative learning, the gulf between the nervous function of a snail and that of a human being is wide, as I said at the outset. To begin with, there is a vast difference in capacity to resolve sensory information and to associate stimuli. Lower species can learn to associate only a narrow range of stimuli; human beings can associate whatever they perceive. Human consciousness makes possible willful forgetting or repression; it places perceptions in a rich emotional context, with gradations of positive and negative feelings that invest associations with value and meaning.

For all of that, much of the biophysics and biochemistry of learning may be quite similar in snails and in human beings. It may be the neural circuitry that makes the difference. Recall that in *Hermissenda* the Type *B* photoreceptors become more excitable in the course of conditioning because fewer potassium channels and more calcium channels open in response to a stimulus—because, in turn, of the increased activation of an enzyme by calcium ions. The Type *A* photoreceptors have the same kind of channels, however, and the same potential for undergoing changes in excitability. It is the wiring diagram of the visual and vestibular pathways that makes the Type *B* cells receive more excitation and so become more excitable when the snail is trained with its head toward the center of rotation. The same wiring diagram determines that the Type *A* receptors, not the Type *B* cells, become more excitable when the animal is trained in the opposite orientation. It is the impact of a particular combination of stimuli, with a particular temporal relation, on a particular genetically programmed wiring diagram that excites one set of cells and inhibits another set.

Consider that billions of neurons in the human brain may have the same potential as the *Hermissenda* cells to become more excitable or more inhibited. It is not an unreasonable assumption; similar potassium and calcium channels have been characterized in invertebrates and in higher vertebrates. Consider also that the relative amount of excitation received by each neuron in a subpopulation of human brain cells must depend on a preexisting wiring diagram and on the nature and timing of sensory stimuli. Just as the relative excitability of photoreceptors is altered when *Hermissenda* learns, so the relative excitability of neurons in the human brain may undergo long-lasting shifts determined by conscious experiences. The fundamental biophysical and biochemical changes responsible for storing learned information need not be very different in human beings from the mechanisms now being revealed for *Hermissenda*.

On the other hand, just what can be learned by an animal, and how much, is structured and limited by the wiring diagram of its neural system. The enormous number of neurons in the human brain and the complexity of its circuitry may provide the potential for recording subtle differences among stimuli, for associating stimuli and for associating the associations, that is, for learning by progressive abstraction. The difference between what people can learn and what a snail can learn may result from the difference in wiring diagrams rather than from any special properties of the cell membrane or any special biochemical control mechanisms that uniquely distinguish human neurons from those of a snail.



POSITIVE-FEEDBACK LOOP accounts for the increased excitability of the Type *B* cell's membrane. At the biophysical level (*black boxes*) increased depolarization during conditioning increases the calcium flux; the increase in intracellular calcium reduces the flux of potassium, further depolarizing the membrane. Two biochemical regulatory steps are shown (*colored cycle*). The high calcium concentration activates a calcium-dependent protein kinase; the enzyme phosphorylates proteins that regulate the membrane channels, reducing potassium flux.



Ottawa, Ontario/Canada

Room with a view

It's called a château and looks like a castle, but it's actually a hotel and please be our guest. Canada's capital abounds with far-away feelings from double-decker bus tours and sedate canal cruises to Changing the Guard on Parliament Hill. The splendours of our heritage are yours to discover in Ottawa, Ontario/Canada. Treat yourself royally! Call TOLL FREE 1-800-828-8585 or from New York State 1-800-462-8404 or write: Ontario Travel, Queen's Park, Toronto M7A 2E5, Ontario/Canada.

ONTARIO
yours
to discover!

Microelectronic Packaging

Methods of housing, cooling and interconnecting the silicon chips in a digital computer have an important bearing on the machine's performance. The aim is to fit the most chips in the least volume

by Albert J. Blodgett, Jr.

The packaging of the microelectronic silicon chips in a high-speed digital computer might seem only incidental to the machine's design. One could argue that the actual processing of information is done entirely by the circuitry on the chips themselves; the functions of the packaging are simply to interconnect the chips and other devices, to distribute electric power and to provide a suitable operating environment. Clearly, every computer needs packaging of some kind, but the nature of the packaging would not appear to have much influence on the functioning of the machine. The facts are otherwise. In many high-speed data-processing units packaging technology is the factor that determines or limits performance, cost and reliability.

One reason packaging has become so important is the imperative to make the central elements of a computing system exceedingly compact. Improvements in the design and fabrication of microelectronic devices have greatly increased the number of logic functions that can be put on a chip as well as the speed at which arithmetic operations are performed. As a result a major source of delay in the central processing unit of many computers now is the time needed for a signal to pass from one chip to another. In order to reduce the delay the chips must be put closer together. Fitting many chips, each with many terminals, into a small volume challenges packaging technology in several ways. First, there is little space available for the thousands of conductors needed to distribute electric power and information-bearing signals among the chips. Furthermore, the electrical properties of this network of conductors must be designed to minimize the distortion of signals, which becomes more difficult as switching speeds are increased and dimensions are reduced. Finally, a dense array of chips gives off a fair amount of heat, which must somehow be removed if the circuits are to operate properly. In many instances the thermal problem is the most challenging one.

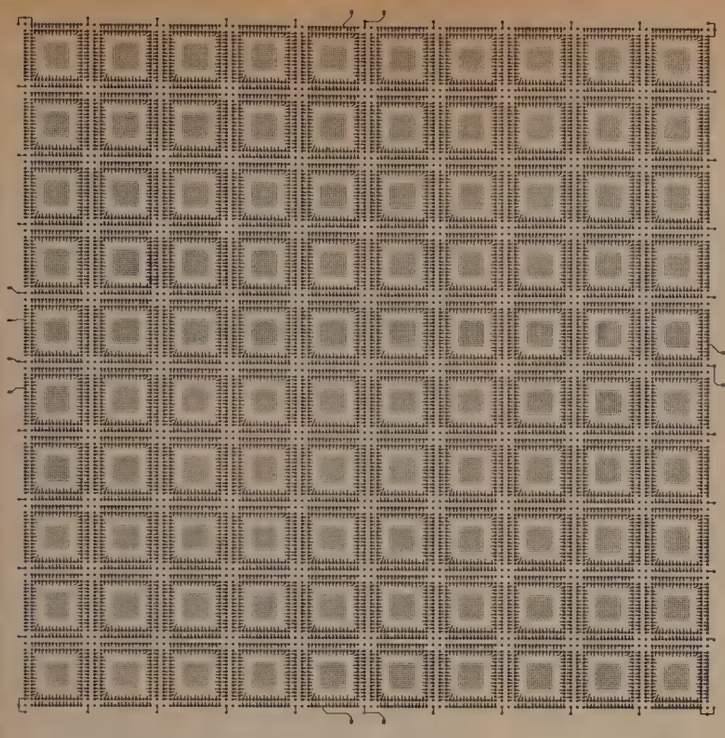
The designers of high-performance digital computers have devised a number of ingenious and quite varied solutions to these problems. Here I shall discuss some general considerations that go into the design of any packaging technology. I shall also describe in some detail the system developed for one high-performance system, the model 3081 processor of the International Business Machines Corporation. It is large "mainframe" computers of this kind that make the severest demands on packaging methods and materials. The technology I shall describe is capable of interconnecting more than 100 chips—which might incorporate circuitry equivalent to the entire central processing unit of an earlier computer—in a module that can be held in one hand.

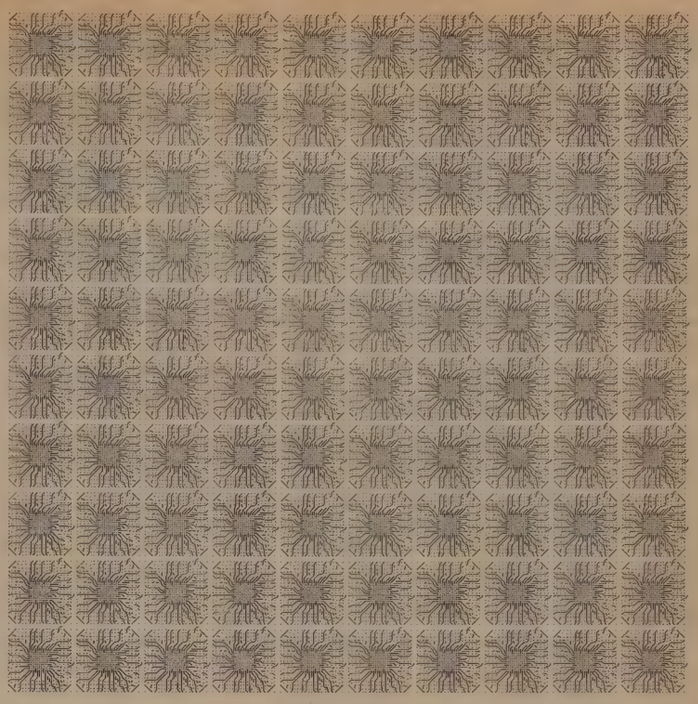
For more than a decade the hierarchy of packages in a typical mainframe computer has had three levels. Each chip is permanently mounted on an individual chip carrier, a plastic or ceramic housing with a dozen or more metal leads. A number of chip carriers, along with a few discrete components such as resistors and capacitors, are mounted on a card made of fiberglass impregnated with an epoxy resin. Metal conduction paths are printed on the surface of the card, and in some cases the card is a laminate with multiple layers of printed wiring. Several cards are mounted through connectors along one edge to a larger board whose printed circuitry establishes connections between the cards. The entire assembly of chip carriers, cards and main board is tied to the rest of the system (including other assemblies of the same kind) by multiconductor cables.

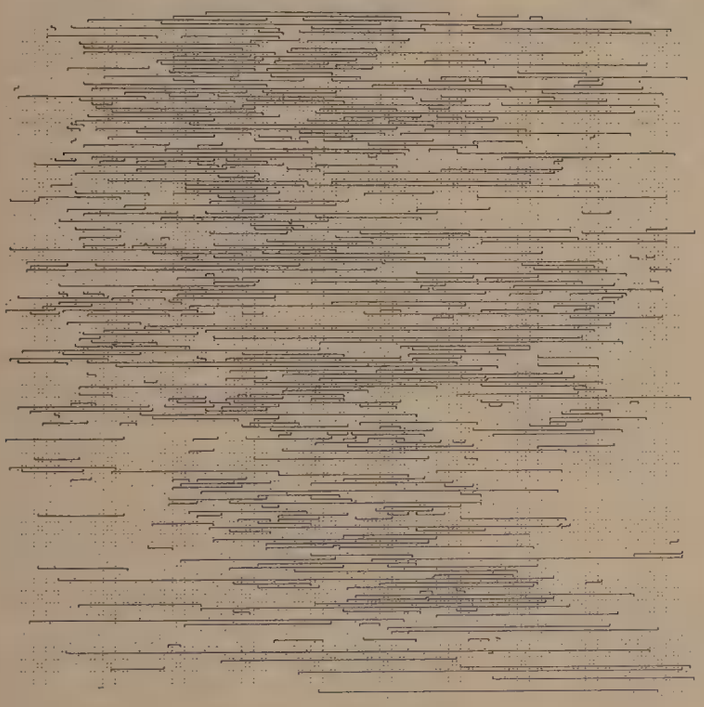
The three-level hierarchy of packages is still the standard in many computers, most notably the ones called minicomputers. In both the lowest- and the highest-speed systems, however, technological advances have made it possible to dispense with one level in the hierarchy. In the low-performance realm of the microcomputer each chip is still housed in a separate carrier; owing to the dramatic increase in the number of circuits per chip, however, the number of chips has been reduced to the point where they all fit on one board. In some of the latest large mainframe computers another approach has been taken: many chips are bonded to a single substrate, forming a multichip module that plugs directly into a large board. Thus in both cases the card level is eliminated.

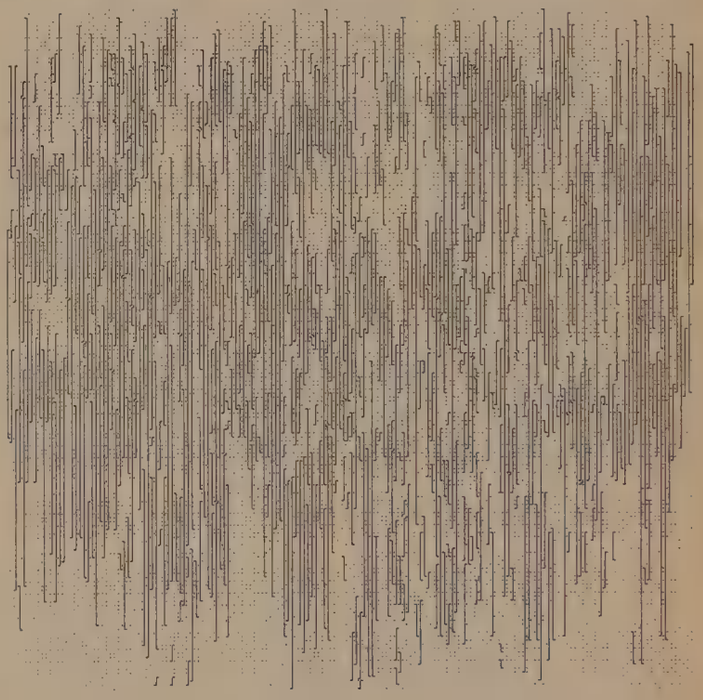
The evolution of microelectronic devices and that of packaging technology are strongly interdependent. For example, putting more circuits on a single chip generally requires adding more electrical contacts, but that is feasible only if the package can accommodate additional connections and signal paths. Similarly, chips with more circuits as a rule consume more power; they can be incorporated into a system only if the package can safely remove the addition-

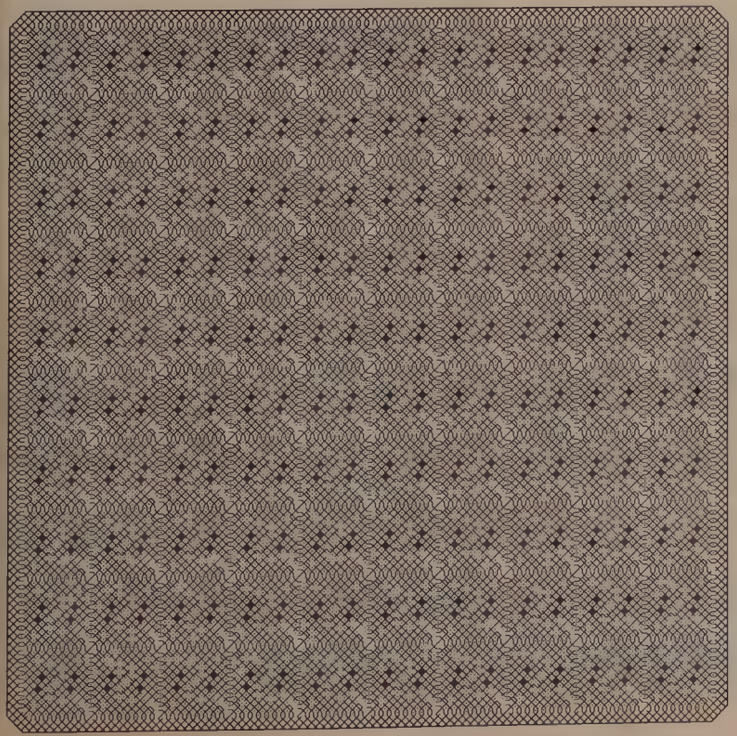
LACEWORK OF CONDUCTORS interconnects 100 silicon chips in the central processing unit of a "mainframe" computer. Each of the patterns on the opposite page forms one layer of a ceramic substrate that can have as many as 33 layers. At the upper left is the top layer, to which the chips are bonded, establishing electrical connections through a dense array of contact pads in the middle of each chip site. At the upper right is one of five redistribution layers, which carry signals from the chip contact pads to engineering-change pads that surround the chip site. The two layers in the middle are signal planes that provide for communication between chips; on a given plane all the conductors are oriented parallel to either the *x* or the *y* axis. A complete substrate has 16 signal planes. Between each pair of signal planes is a voltage-reference plane, such as the one at the lower left, that helps to define the electrical properties of the signal lines. The layer at the lower right is one of three that distribute power to the chips. In the finished substrate the layers are stacked, laminated and fired, so that they fuse into a tilelike ceramic. The patterns are shown at their approximate actual size after firing.

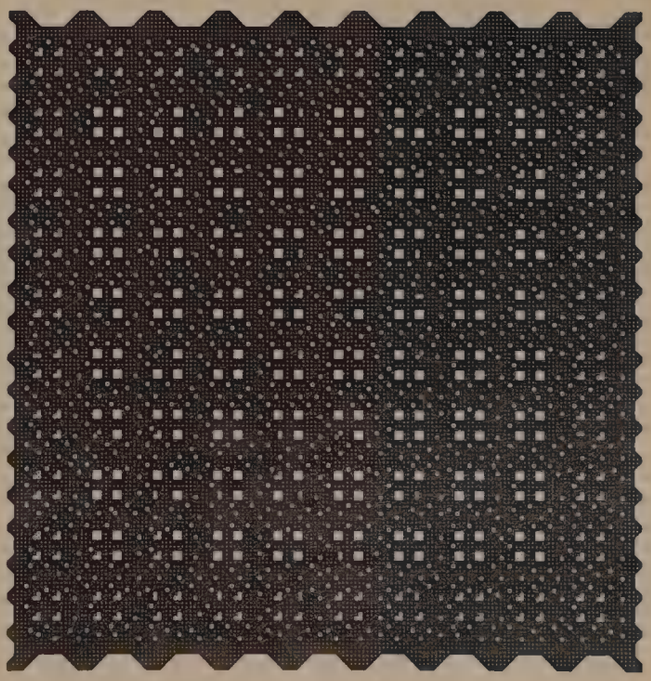












al heat. Because of these relations the constraints that apply to packaging design today can be fully understood only by first considering the nature of the chips that go into the packages.

The crucial event in the evolution of semiconductor technology was the introduction of the integrated circuit, which combines a number of transistors and other basic circuit elements on a single piece of silicon. The first integrated circuits, in the early 1960's, included only from six to eight transistors, diodes and resistors, enough to implement a "gate," or simple logic function. A typical gate receives two or three input signals and issues a single output; the pattern of the input signals determines whether the output is on or off. Several hundred gates might be needed to build a functional unit of a computer such as an adder.

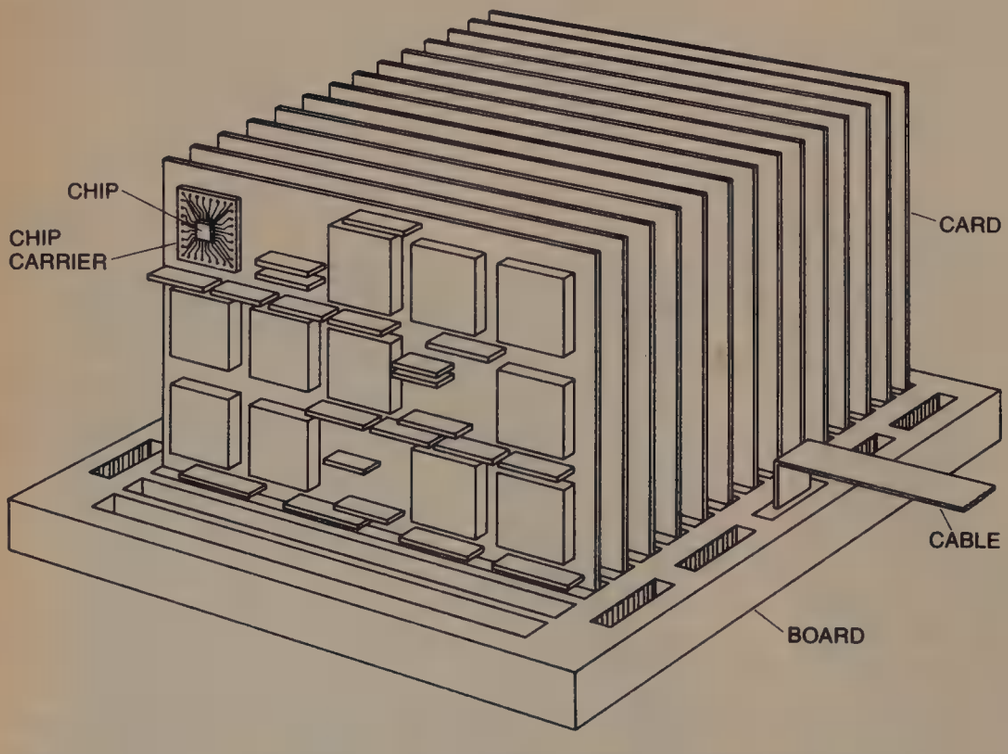
The early integrated circuits were built on a silicon chip with a surface area of a few square millimeters, and the smallest features in the pattern that defined the circuitry were from 10 to 20 micrometers across. The chip communicated with the rest of the system through eight or 10 terminals. The switching delay, which represents the time from the arrival of an input signal to the production of an output signal, was generally between 20 and 40 nanoseconds, or billionths of a second.

Advances in the art of semiconductor manufacture over the past 20 years have been dramatic. The maximum available surface area on a chip has increased by more than 10 times (to about 50 square millimeters), and the minimum feature size has decreased by a similar factor (to less than 1.5 micrometers). High-speed logic chips have been built with more than 40,000 transistors and other circuit elements, organized into 5,000 gates that switch in about two nanoseconds. A chip of this kind can have 200 electrical contacts.

The integrated circuits employed in a high-performance mainframe computer are built out of basic circuit elements called bipolar transistors. Chips based on another kind of transistor, the field-effect transistor, can fit many more circuit elements in a given area. For example, microprocessors and related devices have been built on a single chip with more than 100,000 field-effect transistors. Why, then, do mainframe computers employ bipolar transistors? The reason is that they are faster: as a rule they have a shorter switching delay.

The chips with the highest transistor count are random-access memory arrays, whose function is to store information. Early semiconductor memory devices, introduced in 1971, had 128 cells; each cell was capable of holding one bit (binary digit) of information. Today memory chips with a capacity of 256,000 bits are in production.

For the packaging designer the logic chips of the central processor present a greater challenge than memory chips. In general logic chips dissipate more power and hence give off more heat; they also require more electrical contacts for signal communication. Similarly, a chip based on bipolar transistors consumes more power than a chip with equivalent functions made with field-effect transistors. It follows that packaging design is most critical in the central processor of a high-speed mainframe computer.



HIERARCHY OF PACKAGES in a large digital computer has traditionally had three levels. Each chip is housed in an individual carrier; a number of carriers and other components are attached to a printed-circuit card; several cards are in turn mounted on a larger printed-circuit board. The assembly is connected to the rest of the system by cables. With the recent introduction of a module capable of holding more than 100 chips it has been possible to eliminate the card level and thereby to reduce the number of interconnections and the total wiring length.

Packaging begins where the chip ends: at the metallic pads on the surface of the chip where contact is made with the external wiring. The length of that wiring is among the packaging designer's primary concerns.

In the analysis of many electrical circuits it is reasonable to assume that a voltage applied at one end of a wire appears simultaneously at all points along the length of the wire. Actually the speed of a voltage signal along a wire is finite; it is determined by a property of the insulator surrounding the wire, namely its relative dielectric constant. If the insulator is air (or a vacuum), the relative dielectric constant is 1 and the signal travels with the speed of light in free space, which in units appropriate to this discussion is about 30 centimeters per nanosecond. In other insulators the dielectric constant is larger and the speed is reduced by a factor proportional to the square root of the dielectric constant.

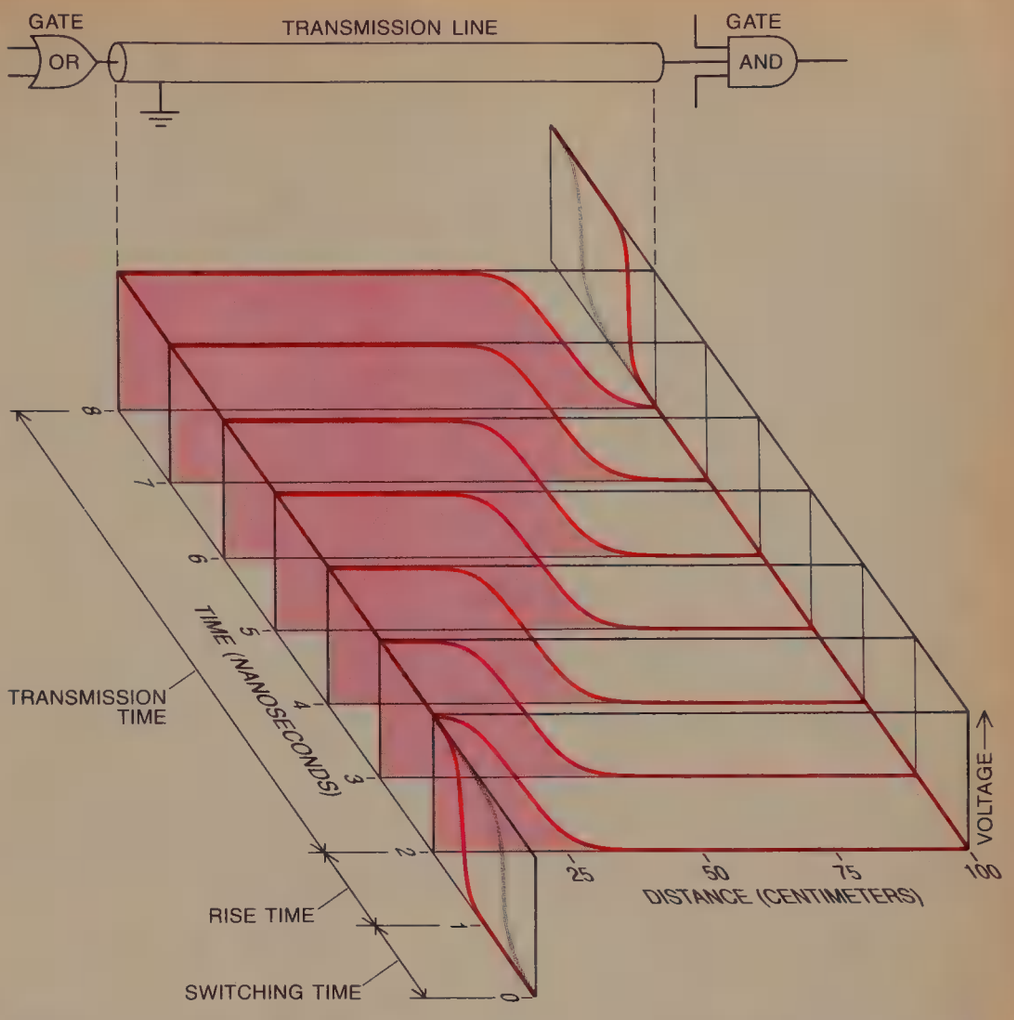
For a fiberglass printed-circuit board the dielectric constant is approximately 4, and so the propagation speed is reduced by a factor of two; in other words, signals travel through the conductors in the board at about 15 centimeters per nanosecond. Because a signal may have to go appreciably farther than 15 centimeters to get from one chip to another, propagation delays can exceed a nanosecond. In slower digital devices a delay of this magnitude is insignificant because the switching delays of the logic gates are tens or hundreds of nanoseconds. In a computer built out of devices that switch in a nanosecond, however, propagation delays clearly have a major influence on the overall speed of operations. It is for this reason that minimizing wire lengths and maximizing circuit density are of critical importance in the design of packaging.

When propagation delays are important, the wire that carries a signal cannot be considered a simple conductor but instead must be treated as a transmission line. The signal is represented as a wave propagating along the transmission line, and the voltage at any point along the conductor depends on both distance from the source and time since the signal was emitted.

In a transmission line the electrical resistance of the conductor is not the only property that affects the propagation of a signal. It is also important to know the inductance, which determines the amount of energy stored in the magnetic field set up by a passing current, and the

capacitance, which determines the energy stored in the corresponding electric field. The inductance and the capacitance depend on the geometry of the transmission line and on electrical and magnetic properties of the materials it is made from; together they define the impedance of the line. For a low-resistance transmission line the impedance is equal to the square root of the ratio of the inductance per unit length to the capacitance per unit length. It is measured in ohms, the same unit employed for resistance, but its effects on a propagating signal are more complicated than the effect of resistance on a steady current.

One characteristic of all waves is that they can be reflected. Similarly, a digital signal can be partially reflected from a discontinuity in the transmission line or from the end of the line. The reflection coefficient, which gives the fraction of the signal reflected, is determined by the impedance and by the load resistance that terminates the line. Suppose a given transmission line has an impedance of 100 ohms. If the load resistance is also 100 ohms, the signal is totally absorbed by the load and none of it is reflected back into the line; this is the ideal situation. If the load resistance is 200 ohms, however, a third of the signal is reflected and adds to the initial signal on the line. A load resistance of 50 ohms also yields a reflection coefficient of one-third, but the reflected signal is subtracted from the initial one. It is clear that such reflections must be controlled to prevent switching errors.



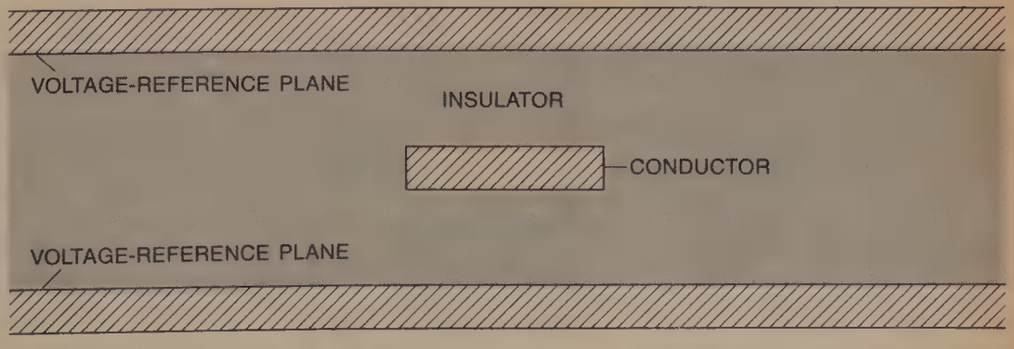
SIGNAL DELAY in a digital computer limits the maximum speed of operation. The delay has three components: the rise time of a voltage applied to a logic gate, the switching time of the gate itself and the transmission time needed for the signal to reach the next gate. When the signal must pass from one chip to another, the transmission delay, which is determined by the packaging technology, is often the longest of the three. Here the rise time and the switching time are each about a nanosecond, but transmitting the signal over a distance of roughly a meter takes six nanoseconds. The signal propagates as an electromagnetic wave in a transmission line; the wave is represented by a graph of voltage as a function of time and distance.

Reflections are only one of several ways the electrical design of a package can modify a signal or inject "noise" into a circuit. For example, two adjacent conductors can be coupled through their mutual inductance and capacitance, so that a signal sent down one line may also appear on the other. Such "crosstalk" must be avoided if the behavior of the system is to be predictable.

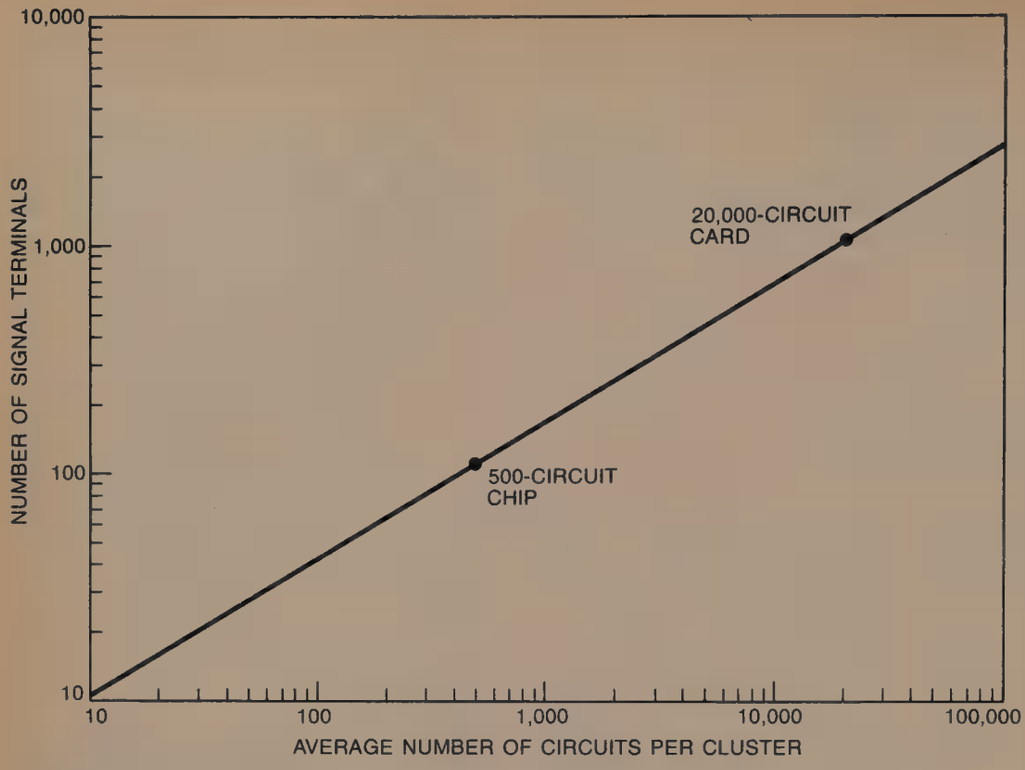
In a high-performance package the basic method of controlling the characteristics of the transmission lines is to separate layers of signal wires with conductive sheets called voltage-reference planes. The reference planes can also provide a path for return currents. Each plane is at a uniform electric potential, either zero volts (ground voltage) or one of the supply voltages needed by the chips and other components. (Hence the planes can also be employed to distribute power.) A signal line encased in an insulating medium and sandwiched between two such planes makes a transmission line whose properties can be calculated. The planes give the line a uniform and well-defined impedance and also inhibit crosstalk between lines in adjacent layers.

The design of a transmission line begins with the specification of its direct-current resistance. The resistance must be small compared with the load resistance or the input voltage will be seriously attenuated when it reaches the output of the line. The resistance per unit length is determined by the resistivity of the material and the cross-sectional area of the conductor; once the material is chosen only the latter property can be altered by the designer. For printed circuits and conductive traces fabricated by similar techniques the cross section is a flattened rectangle.

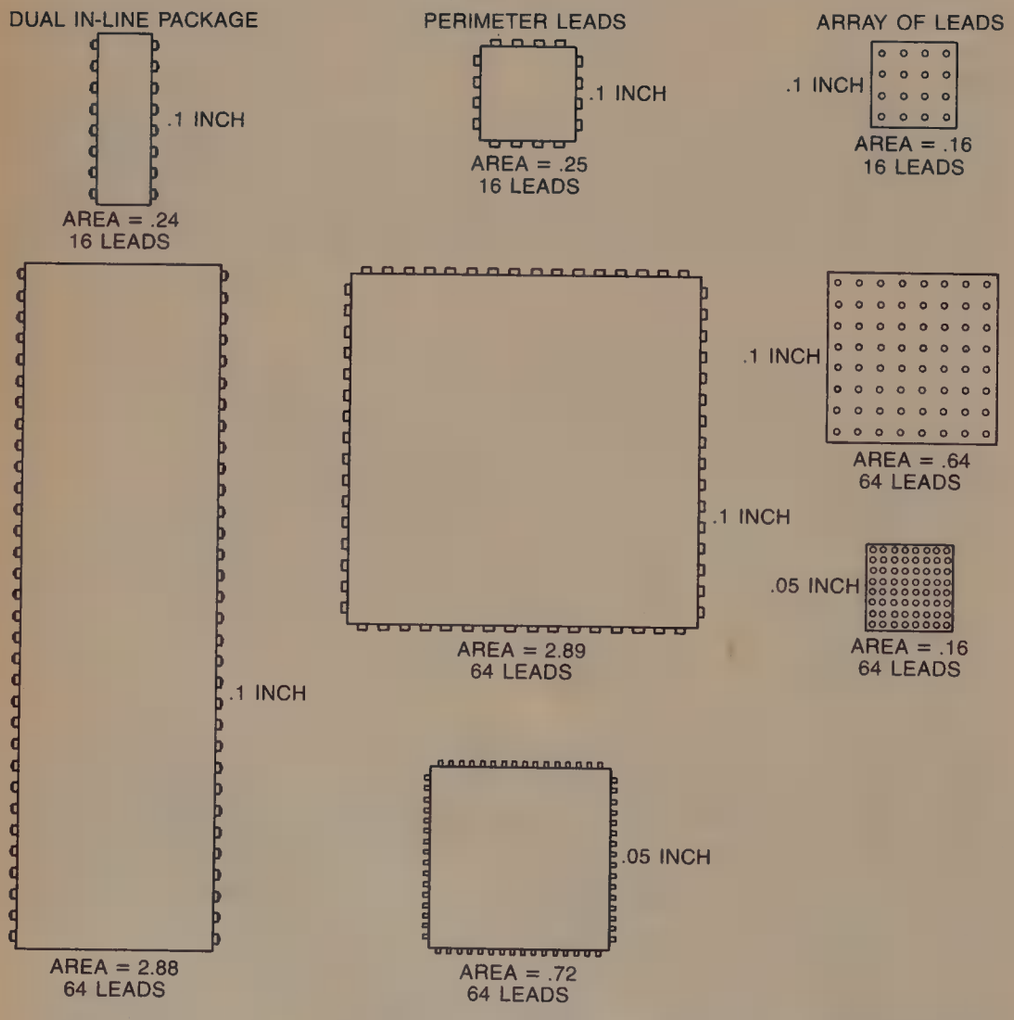
Given the dimensions of the conductor, the line impedance is determined by two additional factors: the dielectric constant of the insulating medium and the distance between the voltage-reference planes. For a particular insulating material the distance between reference planes is adjusted to achieve the desired impedance. The design value depends



TRANSMISSION LINE consists of a conductor embedded in an insulator and sandwiched between two conductive planes. The electrical characteristics of the line depend on the dimensions and the physical properties of the component parts. For example, the speed of a wave propagating through the line is determined by the dielectric constant of the insulator. The parallel conductive planes reduce coupling, or crosstalk, between lines in adjacent layers.



NUMBER OF TERMINALS needed for communication with a cluster of logic circuits can be estimated from an empirical relation called Rent's rule. In a mainframe computer a cluster of C logic circuits needs $2.5C^{.61}$ signal terminals. Thus a chip with 500 circuits would need about 110 terminals, and a card with 20,000 circuits would need about 1,000 signal terminals.



"FOOTPRINT" OF A CHIP CARRIER, or the area it occupies on a card or board, depends on the spacing and arrangement of the terminals. The dual in-line package, with two rows of pins .1 inch apart, is efficient for a chip with 16 leads but not for one with 64. A square package with leads at the perimeter saves no space if the distance between terminals remains .1 inch, but it covers only one-fourth the area when the leads are on .05-inch centers. The most efficient practical configuration for a chip carrier is a full grid of terminals. Indeed, a square carrier with 64 leads on a .05-inch grid takes up less space than a 16-pin dual in-line package does.

on many factors, including the electrical properties of the chips in the system, the dimensions and other specifications of the package and the amount of power available to drive the transmission lines. Typically the impedance is in the range from 50 to 100 ohms.

A conductor between two reference planes can only approximate an actual transmission line. In practice a signal line connecting two chips may follow a tortuous route, threading from one layer of wiring to another. At transitions between packaging levels—such as those from the chip to the carrier or from the card to the board—the electrical properties of the wiring depart significantly from the ideal. As noted above, such discontinuities can cause reflections. They also introduce additional delays, proportional to their capacitance and inductance. The extra delays must be added to the basic propagation delay of the circuit to determine the total delay introduced by the packaging system.

A single transmission line extending between two isolated terminals could be given precisely defined characteristics. The actual task of the packaging designer, however, is to create a network of intermeshed pathways to interconnect thousands of device terminals. The topological complexity of the network is formidable.

The number of signal terminals needed to establish connections at any given level in the packaging hierarchy can be estimated from an empirical relation called Rent's rule, developed in 1960 by E. F. Rent of the International Business Machines Corporation. The rule, which can be applied to chips, multichip modules, cards or boards, gives the approximate number of terminals as a function of the number of logic circuits, C. A form of the rule based on data obtained by analyzing large data-processing systems indicates that the average number of terminals is equal to $2.5C^{.61}$. Thus the number of terminals needed to support 100 circuits is about 40 and the number needed for 1,000 circuits is almost 170.

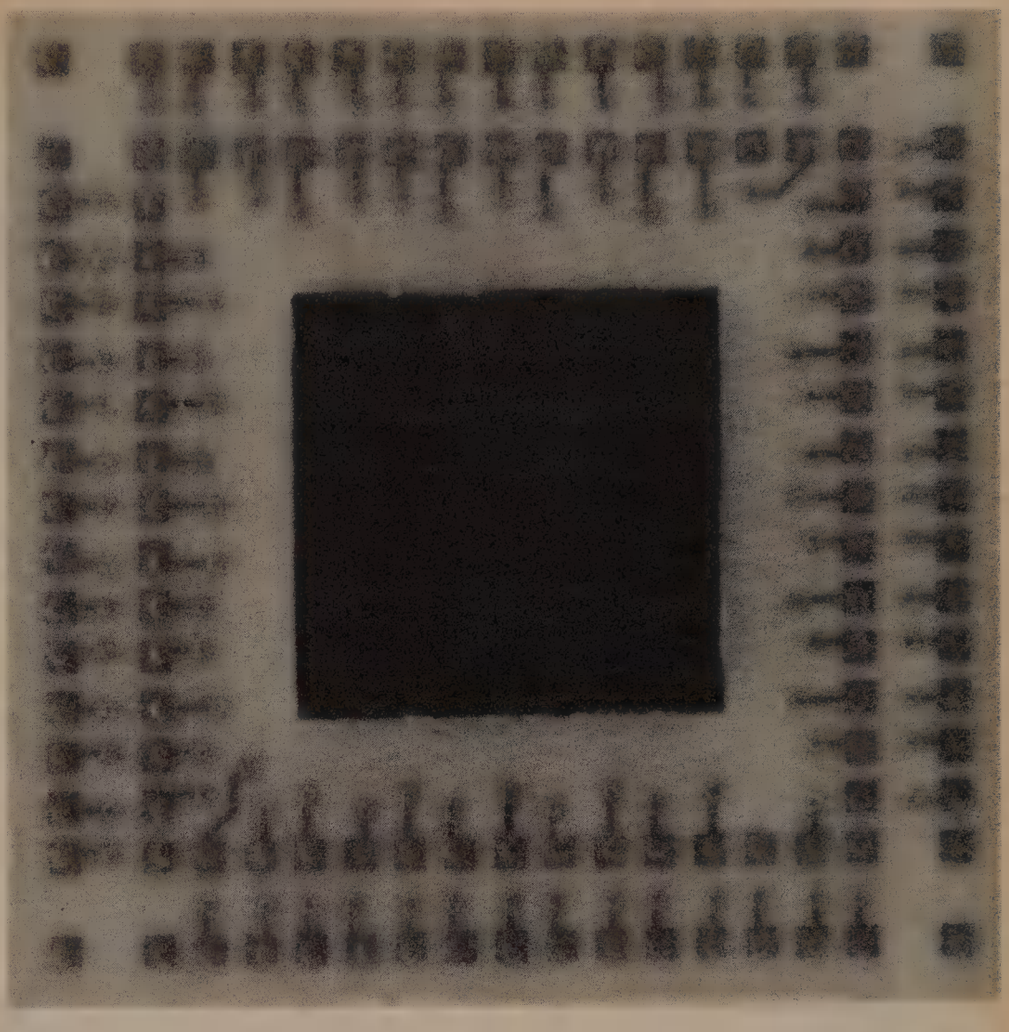
There are certain limitations to the scope of Rent's rule. First, it can be applied only to the logic elements of a processing system; arrays of memory cells need far fewer terminals. Second, each cluster of circuits must be a small, "random" subset of the entire logic complex. If a cluster constitutes a complete functional unit of the computer, then again fewer terminals suffice. Third, it is assumed that information passed between packages is not specially encoded for serial transfer, a technique that can reduce the number of terminals, although only at the cost of slower operation. The clusters of logic circuits for the mainframe computers being considered here meet these conditions.

Consider the packaging of a function-

al unit with 10,000 logic circuits, built up out of chips that have a maximum of 25 circuits each. The chips are housed in carriers with 14 signal pins and mounted on boards with 100 signal terminals. At first it might seem that a 10,000-circuit unit could be assembled out of 400 of the 25-circuit chips. Rent's rule shows, however, that a package with 14 signal pins can support an average of only 17 circuits; hence the actual number of chips needed is 10,000 divided by 17, or 588. (Almost a third of the capacity of the chips remains idle, suggesting that a chip carrier with more than 14 signal pins would be more efficient.)

A second application of Rent's rule indicates that the 100 signal terminals on each board provide enough communications capacity for 424 circuits. The number of boards needed is therefore 10,000 divided by 424, or 24, and each board can hold an average of about 25 chip carriers.

Rent's rule estimates only the number of packages needed at a given level in the hierarchy; the actual size of a card or board must be determined from the details of the physical design. Much depends on the lowest-level package, which is generally the single-chip carrier. The factors that need to be considered in card design include the configuration of the terminals on the carrier, the maximum density of the signal paths, the electrical performance required, the power requirements of the chips and the cooling capacity of the overall packaging design.



MULTICHIP CARRIER can greatly reduce the average distance between chips and hence the average transmission delay. Here a single chip is shown mounted face down on a multilayer ceramic substrate that can accommodate about 100 chips. Both power and signals pass through conductors buried within the substrate. The double row of pads surrounding the chip gives access to the signals, so that the assembly can be tested and connections can be altered.

A chip carrier in its simplest form is a space transformer. It serves as a bridge between the small and closely spaced contact pads on the surface of the chip and the larger wiring network on the card or board. One of the commonest chip carriers is the dual in-line package, a rectangular enclosure of plastic or ceramic with a row of leads along each of its two longer sides. The leads are arranged on .1-inch centers and are inserted into holes in the printed-circuit board. The dual in-line package makes fairly efficient use of space on the card for chips that have comparatively few terminals (say from 10 to 20), but it becomes unwieldy as the number of leads increases. For a 64-pin chip not only is the package greatly lengthened but also its width must be increased to allow space inside the package for connections between the leads and the chip contact pads.

The industry standard is now shifting from the dual in-line package to a square chip carrier with leads, or contacts, on all four sides. The leads are on .05-inch centers and are soldered to corresponding pads on the surface of the printed-circuit card. The design significantly reduces space requirements. A square chip carrier with 64 perimeter leads on .05-inch centers occupies only a fourth the area taken up by a 64-pin dual in-line package.

The densest terminal configuration for a chip carrier is a full array of pins or contact pads. Sixty-four terminals in an eight-by-eight array on .05-inch centers take up less than a fourth the space needed for the same number of leads in a square perimeter arrangement at the same spacing. The advantage of the array format becomes even greater as the number of terminals increases. Making connections to the denser array of pins or contacts, however, requires a printed-circuit board that is more complex and hence more expensive.

As the number of terminals per chip and the number of chips per unit area increase, the major limiting factor becomes the capability of the board or multichip module to provide interconnections. The simplest printed-circuit boards have only one layer of signal lines, and so the wiring must be laid out with careful attention to an obvious constraint: no two conductors can be allowed to cross. Multiple layers of wiring can eliminate this problem, particularly if all the conductors in a given layer are oriented along one major axis, either x or y. On the other hand, with a multilayer board space is needed for vias, or vertical pathways, that connect the layers. A signal crossing the board diagonally would be directed through a via to an x plane, where a signal line would take it along one edge of the board, through another via to a y plane, with a signal line perpendicular to the first one, and finally through a third via to the surface of the board at the destination.

The layout of a multilayer board is governed by design rules that specify the dimensions of the conductors and where they can be placed. For example, the vias are generally arranged in a grid; not every intersection of the grid lines necessarily has a via, but vias can be installed only at the grid points. The signal lines must run between the rows of vias. Inserting multiple conductors between adjacent vias can increase the overall wiring density, but it also complicates the design of the board since a conductor may be blocked from a particular via by another conductor.

The design rules establish the maximum wiring capacity of a packaging technology. The maximum length of wiring per unit area of board surface is equal to the number of signal layers

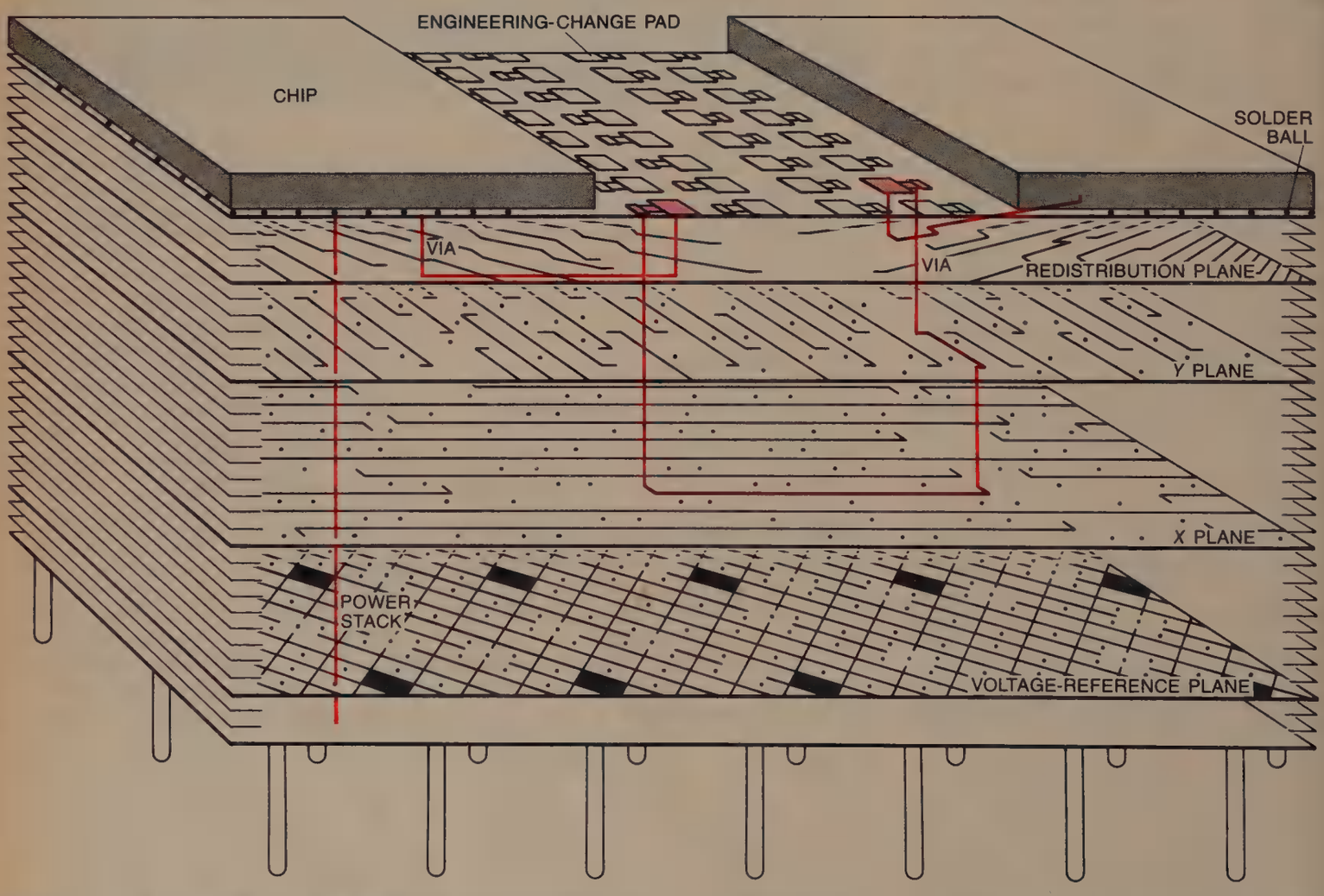
multiplied by the number of signal lines between adjacent vias and divided by the grid spacing of the vias. For example, a board with two signal planes, two wires between vias and vias on .1-inch centers has a wiring capacity of 40 inches per square inch. Both the wiring capacity and the efficiency with which it can be put to use are strongly influenced by the via size and the availability of vias for communication between signal planes.

A technology that packs many chips into a small volume is futile if the heat generated by the chips cannot be safely carried off. The property that determines the cooling capacity of a chip carrier or module is its thermal resistance, which is closely analogous to electrical resistance. According to Ohm's law, the resistance between two points in an electrical circuit is equal to the voltage difference between the points divided by the current flowing between them. Similarly, the thermal resistance of a package is the temperature difference between the heat source (the chip) and the ultimate heat sink (the ambient air) divided by the heat flux passing through the package. When the system is in a steady state, the heat flux must be equal to the power being dissipated by the chip. The thermal resistance can be expressed in units of degrees Celsius per watt.

The thermal path from the chip to the air can be considered in two parts. First, heat is conveyed, usually by conduction, from the semiconductor junctions—the areas within the structure of a chip where most of the heat is generated—to the surface of the package. The thermal resistance of this part of the path is called the internal resistance and depends mainly on the geometry of the chip carrier and the thermal conductivity of the materials. Second, heat is removed from the package itself, in most cases by moving air; the mechanism of heat transfer is either forced or natural convection. The external thermal resistance is a complicated function of many factors, including the area and thermal emissivity of the module and the velocity and turbulence of the airstream.

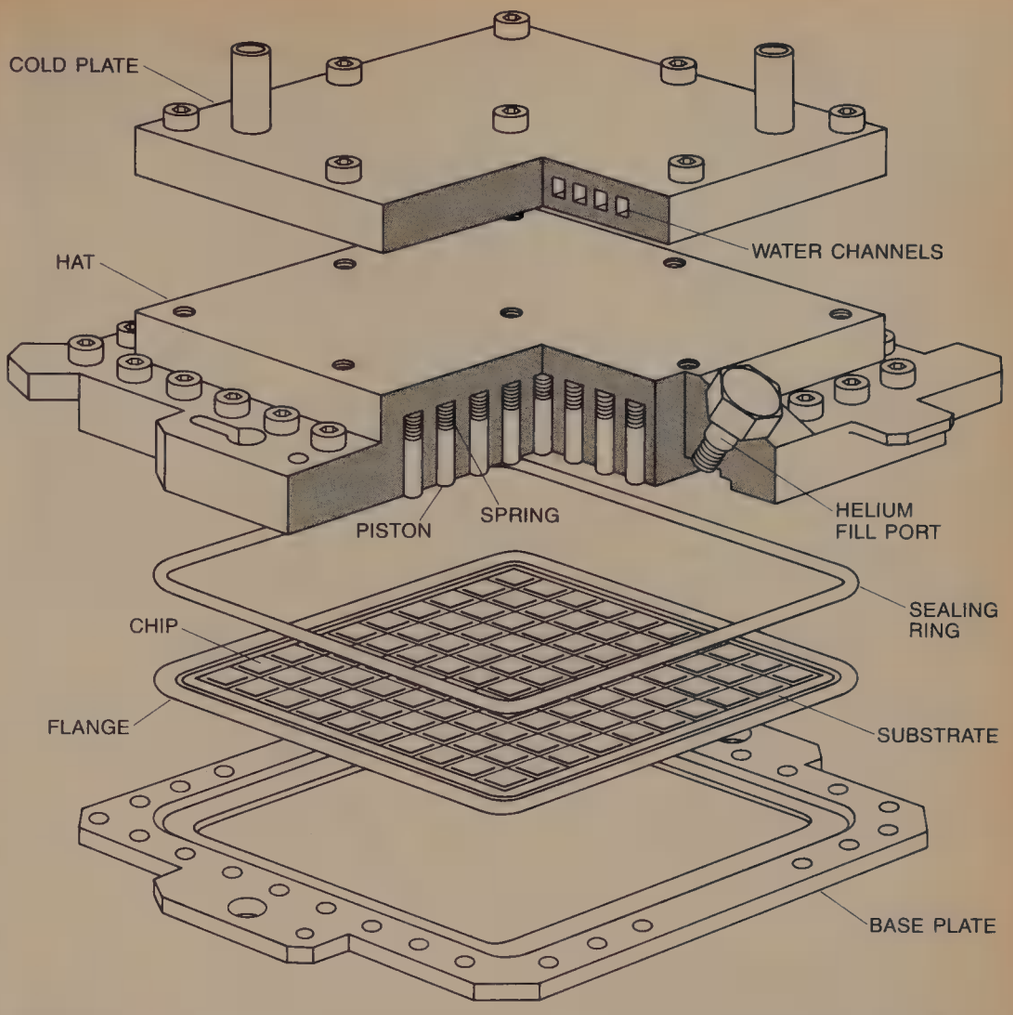
The overall thermal resistance of a typical plastic dual in-line package in moving air is about 50 degrees C. per watt. Thus a chip dissipating .5 watt would undergo a 25-degree rise in temperature. The internal resistance can be reduced by replacing the plastic package with a ceramic one, which has a higher thermal conductivity, or by enlarging the area over which the chip is bonded to the carrier. Such measures can reduce the thermal resistance by half and thereby double the power capacity. The external resistance can also be reduced, for example by adding a heat sink to increase the effective area of the package or by increasing the velocity of the air, but it is very difficult to reduce the overall resistance by more than another factor of two.

For a package with a given thermal resistance the maximum power dissipation depends on the temperature difference available for cooling. Although some silicon devices can operate at a junction temperature exceeding 150 degrees C., the maximum operating temperature of digital circuits is usually



MULTILAYER CERAMIC SUBSTRATE can accommodate 320 centimeters of wiring per square centimeter of surface area. Here a small region of the substrate is shown in a schematic and greatly magnified cross section. A typical signal path proceeds from a chip down a "via" that passes between the layers of the structure. A conductor on one of five redistribution layers carries the signal to another via that returns to the surface at an engineering-change pad. The signal then passes through a metallic link along the surface and into a third via, which takes it to a signal plane deep within the substrate. A conductor oriented along the *x* axis leads to still another via and a *y*-oriented conductor; a final via returns to the surface at another chip site to complete the signal path. Some signal paths go to pins on the bottom of the substrate, where connections are made to other multichip modules. Power is distributed by three planes at the bottom of the substrate. The pattern of connections can be modified by cutting the link at an engineering-change pad and bonding a new wire to the pad.

limited to between 75 and 85 degrees to ensure reliability and uniformity of electrical properties from chip to chip. The inlet temperature of the cooling airstream can be as high as 30 degrees, and the air heats up (by as much as 10 to 15 degrees) as it passes through the package. Hence the maximum available temperature difference is less than 50 degrees. These calculations imply that the power-handling capacity of a ceramic dual in-line package is less than two watts. In general the maximum heat flux with air cooling is about two watts per square centimeter at the chip level and about .5 watt per square centimeter at the module level. (The values can be increased by using special heat sinks and high-velocity chilled air.) These limits represent important constraints on circuit power and circuit density and therefore on performance.



THERMAL-CONDUCTION MODULE houses the multilayer substrate and cools the 100 or more chips mounted on its surface. The substrate is clamped between a baseplate and a "hat" equipped with spring-loaded metal pistons. Each piston presses against the back of a chip, conducting heat to a cold plate bolted on top of the hat. The cold plate in turn gives up the heat to chilled water that is pumped through channels in the plate. The module's ability to dissipate heat is enhanced by filling the internal volume with helium, which has a higher heat conductivity than air. The rated power-handling capacity of the thermal-conduction module is 300 watts.

Some designers of high-speed computing systems have explored a number of alternatives to packaging technology based on the air-cooled chip carrier. Here I shall describe one high-performance technology, which I believe has extended the state of the art. It was developed for a new series of mainframe computers by scientists and engineers at the IBM facilities in East Fishkill, Endicott and Poughkeepsie, N.Y., and in Sindelfingen, West Germany. Two fundamental design objectives were to reduce the number of interconnections between packaging levels and to reduce the total wiring length. These goals offered three potential benefits: higher speed, lower cost and improved reliability. The package that resulted from these efforts has two main components: a multilayer ceramic substrate on which the chips are mounted and through which all interconnections are made, and a module assembly that provides a direct thermal path from the back of the chips to a water-cooled heat sink.

The ceramic substrate compresses a wiring network of extraordinary complexity into a square, tilelike object 90 millimeters on a side and roughly five millimeters thick. On the top surface are sites for between 100 and 133 high-speed chips, with a total of more than 12,000 chip contact pads. On the bottom surface of the substrate are 1,800 pins that supply power to the chips and route signals to and from the next level in the packaging hierarchy. Both the chip terminals and the module terminals are organized as two-dimensional arrays to minimize the space required. The substrate itself has 33 layers of conductors, which are interconnected by more than 350,000 vias.

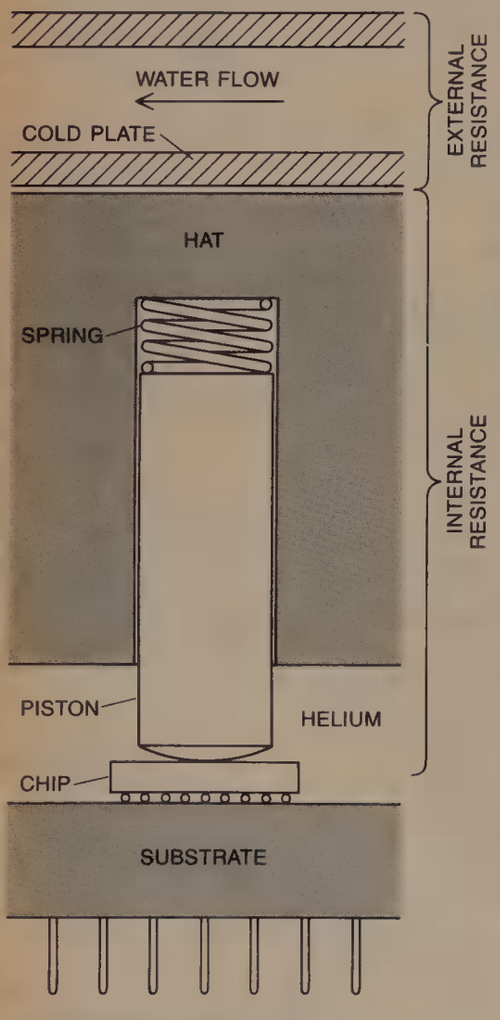
Sixteen of the layers in the substrate are x or y planes of signal wiring. The design rules for these layers allow vias to be placed on grid points with a spacing of .5 millimeter. Just one signal line is allowed between adjacent vias, eliminating the possibility of conflict for access to a via. Thus the 16 signal layers have a maximum wiring capacity of 320 centimeters per square centimeter of substrate. A typical substrate has 130 meters of wires on these planes. A voltage-reference plane is inserted between each pair of x and y layers to control the impedance of the signal lines. The characteristic impedance is 55 ohms.

The top five layers of the substrate have the most vias. Here the vias are on a .25-millimeter grid to match the area array of 120 contact pads on each chip. These layers are employed to redistribute the signal lines (96 per chip site) from the chip pads to a set of contact pads that surround each site. From there the signal lines return to the interior of the substrate. The surface pads allow the module to be tested with the chips in place. Furthermore, if changes in wiring are needed, a small link on the surface can be severed to isolate any signal line from the internal wiring of the substrate. A new connection can then be made by bonding a fine wire to the metal pad and laying it in the channels between the chips. The ability to make such engineering changes is particularly important during the development of a new product. There are also methods for replacing an individual chip because of either a design change or the failure of a component.

Power is distributed by three planes at the bottom of the substrate; two of the planes carry the voltages required by the chips and the third is at ground potential. The planes themselves are supplied in parallel by an array of pins distributed uniformly across the bottom surface; 500 of the 1,800 available pins are used for this purpose. From the power planes the current flows directly to the chip power pads through parallel vertical stacks of vias. The substrate is designed to supply up to four watts to each chip site, although not all the chips draw the maximum current. The module as a whole is limited to a total of 300 watts, or an average of about three watts per chip. The power-supply voltage loss due to resistance in the package is less than 15 millivolts.

As might be expected, some compromises had to be made in matching the

design to the materials for such a complex packaging technology. The principal component of the substrate is alumina, or aluminum oxide, a ceramic that was chosen because of its superior mechanical properties and its ability to withstand the various chemical and thermal processes employed in manufacturing the substrate and module. The principle drawback of alumina is its relatively high dielectric constant of 9.4. As a result the propagation speed of signals in the module is lower than it would be in a fiberglass printed-circuit board; since chips can be mounted directly on the ceramic substrate, however, the average distance between chips is much smaller than it would be with card-on-board technology, and so the total propagation delay is significantly less. Another compromise was made in choosing the material of the conducting paths inside the substrate. Because the ceramic must be fired at 1,500 degrees C., the conductors must be made out of a refractory metal; other metals have higher electrical conductivity, but their melting point is too low. (Copper, for example, melts at 1,083 degrees C.) The metal chosen is molybdenum, whose bulk resistivity is approximately three times that of copper. This constraint is compensated for by the high density of vias in the design; large currents are supplied directly to the chips by stacks of vias.



THERMAL RESISTANCE of the module determines the maximum temperature rise of a chip with a given power dissipation. The resistance of the heat path from the chip through the piston (and the helium atmosphere) to the hat is defined as internal; the resistance from the cold plate to the flowing water is external. The total resistance of the package is some 11 degrees Celsius per watt per chip site. Therefore a chip that dissipates four watts undergoes a rise in temperature of 44 degrees above the water temperature.

The fabrication of the multilayer ceramic substrate begins with the casting of the individual layers. Ceramic and glass powders are mixed with an organic binder and solvent to form a slurry, which has the consistency of paint. The slurry is deposited on a moving plastic belt and passes under a blade that sets the thickness of the layer. A long drying oven drives off the solvent, leaving a cohesive but still flexible material that looks rather like thick paper. Square blanks called green sheets are then cut from the web. ("Green" is the ceramist's term for unfired material; the sheets are actually white.) Alignment holes are punched in the corners of each sheet to aid in subsequent operations.

The next step is the punching of via holes. It is done under computer control by a high-speed multiple-punch-and-die machine. In the top layer of a 100-chip module, for example, every chip site has an identical array of holes. One hundred punches are therefore mounted in a grid with a spacing equal to the spacing of the chip sites. Each operation of the tool punches a single hole at the same relative position in each chip site; the entire sheet is then moved slightly and the next set of 100 holes is punched.

The metal patterns are laid down on the green sheets by a process similar to stenciling or silk-screen printing. A paste of molybdenum in a binder and solvent is extruded through a metal mask that has the pattern of the wiring cut into it. (The pattern itself is generated automatically by data from the computer-aided-design system.) The paste is applied under pressure so that the punched via holes are also filled. The metallized sheets are then dried and inspected. The detection of flaws in individual sheets before they are combined with other sheets to form a complete substrate is important to maintaining the overall yield of the manufacturing process.

The sheets that pass the inspection are stacked in the appropriate sequence and laminated under high pressure at 75 degrees C. Because vias only 120 micrometers in diameter must be continuous from one layer to the next, control of dimensions and alignment is critical. The green laminate is trimmed to size and subjected to a long firing cycle in which the peak temperature of more than 1,500 degrees is reached in a hydrogen atmosphere. At lower temperatures the organic material decomposes and becomes volatile, and at the higher temperatures the ceramic and the metal sinter into a monolithic structure.

The rate of heating must be carefully controlled; if the temperature were to rise too quickly, the organic binder would volatilize faster than it could diffuse to the surface, causing the substrate to delaminate. During the sintering process the substrate shrinks by approximately 17 percent in each dimension, for a total volume reduction of about 40 percent. Given the tight dimensional tolerances of the final assembly, it is clear that the amount of shrinkage must be known precisely when the patterns are first inscribed on the green sheets. After firing, the substrate has the size, shape and characteristic hardness of a ceramic tile; if it is struck, it rings.

The exposed metal areas on both surfaces of the completed substrate are plated with nickel and then with gold. An extensive electrical test is done by an automated tester that again utilizes data from the computer-aided-design system to verify the correct pattern of interconnections. The machine must confirm that each pad is connected to other pads as specified by the design; in addition the machine must make certain there are no extra, improper connections. When the testing is completed, the 1,800 pins are brazed to the bottom surface; a metal flange is attached, also by brazing, in the same operation.

The integrated circuits are attached to the substrate by a method developed at IBM for an earlier generation of computers. First a tin-lead solder is evaporated through a metal mask onto the contact pads on the surface of the chip. The chip is then heated in an inert atmosphere to melt the solder, which, under the influence of surface tension, forms a spherical droplet on each terminal. The solder is allowed to harden again, and the chip is inverted over the substrate with the contact pads aligned. When all the chips are in place, the assembly is once more heated to the melting point of the solder; each solder pad assumes the shape of a truncated sphere, connecting the chip and the substrate terminals electrically but still holding the chip above the surface. The mounted chips can be tested through the surrounding engineering-change pads. Any surface wires needed are then attached by ultrasonic bonding.

The multilayer ceramic substrate, with its flange and pins, forms the base of the thermal-conduction module, which has two other major components, called the hat assembly and the cold plate. The hat is clamped to the substrate flange, compressing a pliable ring with a *C*-shaped cross section and thereby sealing the internal volume. The cold plate is bolted on top of the hat. Within

the hat is an array of spring-loaded aluminum pistons with the same pattern as that of the chips on the substrate. Each piston presses against the back of a chip, conducting the heat upward to the cold plate. The cold plate in turn is cooled by water chilled to 24 degrees C. flowing through internal channels at a rate of 40 cubic centimeters per second. The thermal properties of the module are enhanced further by filling the sealed volume with helium, which at room temperature has a much higher thermal conductivity than air. The helium reduces the internal thermal resistance by more than half.

The assembled module has an internal thermal resistance (measuring from the chip up to the cold plate) of nine degrees per watt per chip site and an external resistance of two degrees per watt per chip site. The module is conservatively specified for use at a maximum of four watts per chip and a total of 300 watts. From the thermal resistance it can be calculated that a chip dissipating four watts should reach a temperature of 68 degrees, which is well below the maximum operating temperature of the circuits. The heat flux is about 20 watts per square centimeter at the chip level and four watts per square centimeter at the module level, an order of magnitude greater than the heat flux in a typical air-cooled package.

The thermal-conduction module represents only one level out of three levels in an advanced-technology central processing unit. The first level is that of the chips themselves, which were developed at the IBM laboratory in East Fishkill. The logic chips all have the same underlying structure of 704 basic logic cells; over this structure, however, there are three levels of metallic conductors that are customized to create various configurations of gates and other devices. The metal layers also supply power to the circuits and distribute signals between the cells and the contact pads. The bipolar logic circuits on the chips have a switching delay of 1.1 nanoseconds.

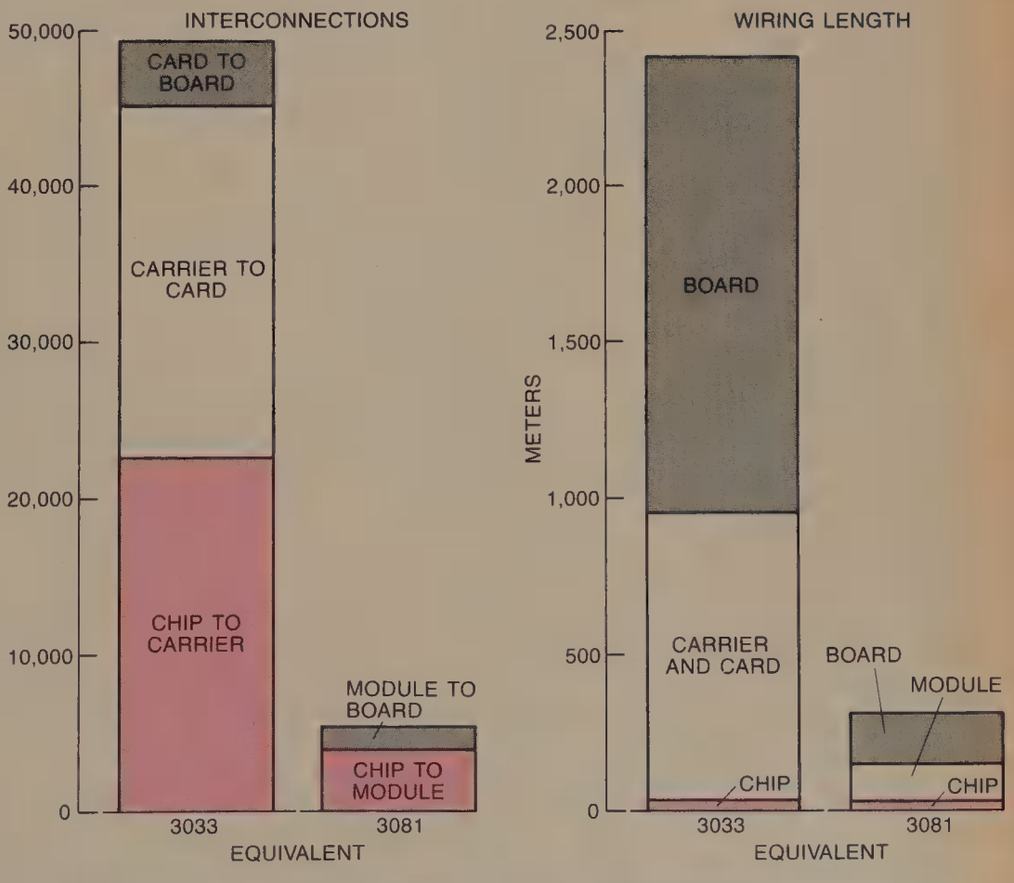
The third component of the system, which represents another major advance in packaging technology, is the large printed-circuit board that interconnects up to nine thermal-conduction modules. The board, developed at IBM's Endicott facility, has 20 layers, including six impedance-controlled signal planes. A nine-module board measures 60 by 70 centimeters and provides connections between the 16,200 pins of the modules and more than 2,000 additional terminals for cables leading to other subassemblies. The board also supplies up to 600 amperes of current to the modules.

The thermal-conduction module and the associated technologies were introduced in 1981 in the IBM 3081 high-



BED OF NAILS on the underside of the thermal-conduction module is made up of 1,800 connecting pins. Five hundred of them supply power to the module; the rest are available for communication with other modules and with other components of the computing system.

performance data processor, developed at the Poughkeepsie laboratory. In a typical module in the 3081, 52 of the chip sites are occupied by logic chips. There are also 34 array chips, a form of high-speed semiconductor memory employed for data and instructions that must be immediately available to the central processing unit, and five terminator chips, which are arrays of resistors used to match the impedance of long, off-chip transmission lines. A module with this complement of chips holds more than 25,000 logic circuits,



EFFECT OF PACKAGING on performance and reliability is suggested by a comparison of the packaging technologies employed in two mainframe computers made by the International Business Machines Corporation. The IBM 3033 is built with individual chip carriers, cards and boards; the 3081 employs the thermal-conduction module. Here the comparison is made between a single thermal-conduction module and a set of 3033 components with the same number of logic circuits. The number of connections between packaging levels in the thermal-conduction module is smaller by a factor of 10; furthermore, most of the remaining connections are solder joints between the chip and the substrate, which are more reliable than mechanical connections. The signal wiring length in the thermal-conduction module is one-eighth that in the 3033 equivalent, leading to a commensurate reduction in signal-transmission time.

65,000 array bits and almost 500 terminating resistors.

It is instructive to compare the packaging technology of the 3081 with that of another high-performance IBM computer, the 3033, which is built with a more conventional packaging technology. In the 3033 each chip has an average of 12 logic circuits and is housed in an individual chip carrier. The carriers in turn are mounted with terminating resistors and other components on printed circuit cards, which connect along one edge to larger circuit boards. If the functions of the typical thermal-conduction module described above were implemented in the 3033 technology, 1,880 chip carriers would be needed for the logic circuits, another 80 chip carriers for the memory arrays and additional components for the line terminators. To mount and interconnect all these components would take 52 printed-circuit cards, four larger boards and the cabling required for connections between the boards. Note that these components would be needed to reproduce the functions of a single module; a full-featured 3081 system has 26 modules.

Three effects of the packaging technology based on the thermal-conduction module are particularly noteworthy. First, the large reduction in packaging hardware results in a significantly lower cost. Second, the system has improved reliability. The main sites of failure in electronic assemblies are the connections between package levels. The thermal-conduction module has eliminated one level of packaging entirely and has reduced the number of logic signal connections between levels by a factor of almost 10. Furthermore, most of the remaining connections are the chip-to-module solder joints, which are inherently more reliable than mechanical connections at higher packaging levels.

The third effect is improved performance. The total length of logic wiring in the 3081 processor is roughly one-eighth what it would be in equivalent 3033 technology. The result is a fourfold reduction in the total packaging delay of the central processing unit, including the modules, the boards and the cables; this in turn allows a twofold decrease in processor cycle time.



PRINTED-CIRCUIT BOARD mounted on a massive steel frame represents the next level (after the thermal-conduction module) in the packaging hierarchy of the 3081 computer. The board accepts nine modules. Here eight modules have been installed and connected to the water-cooling system; the ninth position is left vacant, exposing the connectors that receive the 1,800 pins on each module.



Stretch.

It's the only way to see how far there really is. For thousands of young people, one place to stretch is in local swim clubs and the novice programs of United States Swimming. They get coaching, fun, encouragement, a little taste of competitive swimming and a big lesson that we can all afford to remember. They're learning that nothing is too new to try. And that's a lesson that reaches a long way past the edge of the pool.

PHILLIPS 66

Phillips Petroleum has sponsored United States Swimming for eleven years, and United States Diving for five, helping with operating costs of national and international competitions that lead to the Olympics. We continue to support them. Because their ideals touch the lives of everyone.

The Development of Palm Leaves

The compound leaves of plants usually arise either from differential growth or from selective cell death. Palm leaves, however, follow a developmental pathway that combines both of these processes

by Donald R. Kaplan

One of the central questions of modern biology is: How does a mature plant or animal develop from the fertilized egg and then from the embryo? Pioneer biologists described the stages of development in many organisms and tissues in great detail, but the investigation of the underlying processes that control development had to await the rise of molecular biology. Moreover, even describing the subtle stages of development in many tissues had to await the rise of methods such as electron microscopy. Clearly the underlying processes in the stages of development of a tissue would be difficult to identify without identifying the stages themselves.

A prime example of where a complex developmental pattern has had to be clarified before its control could be investigated is the leaf of one family of flowering plants: the palms (Palmae). All flowering plants have three main structures: roots, stems and leaves. Of these structures leaves are the most diverse. For example, a leaf, in addition to its principal role as an organ of photosynthesis, may be modified into a protective seal over a new bud, into a climbing organ, into the reproductive organs of a flower or even into an insect trap. An unmodified leaf is characterized by a broad, flat green expanse, the part of the leaf that has the greatest concentration of chlorophyll, intercepts sunlight and acts as an organ of gas exchange. This, however, is only one of a leaf's three component parts. It is known formally as the lamina (from the Latin for "plate") and less formally as the blade.

What supports the blade is the second component of the leaf: the petiole (another borrowing from the Latin, this time by the great Linnaeus, that means "little foot"). The petiole serves largely as a conduit, carrying the nutrient products of photosynthesis from the leaf blade to the third leaf component. That component is the leaf base, which joins the other parts of the leaf to the plant stem. As well as providing mechanical support for those parts and conveying nutrients from the petiole to the stem (and thus to the plant as a whole) the base surrounds and protects the rest of the leaf in its younger and more embryonic form, when a leaf-to-be exists only as part of a terminal bud on a shoot.

Exactly how does a leaf develop from the bud? Although I shall be discussing the process primarily in terms of representatives of the palms, it will be useful to include representatives of a not very distantly related family: the arums (Araceae). The palms and the arums grow leaves from the embryonic state to maturity by means of distinctly different pathways that were not fully understood until recently.

Plant development is quite unlike animal development. With most animals the formation of new organs is confined to the earliest phases of embryonic growth. With plants the process is a continuous one. New organs arise from perpetually embryonic growth centers: undifferentiated tissues consisting of cells capable of transformation into a variety of plant organs. The growth centers, at the extremities of the plant, are embryonic tissue at the apex of the root and shoot. Such tissue is termed the apical meristem.

In a typical shoot system the meristem initiates new increments of stem growth, together with leaves along the stem, in a precise geometric pattern. Since new stem increments and leaves arise over a prolonged period, it might seem that a plant shoot would be a set of structurally identical units, like the segments of an earthworm. For example, those who study leaf development have taken for granted that if the leaves below the terminal bud of a shoot were arranged in order of increasing age, the sequence would illustrate the stages a leaf at a given position on the stem would go through if its development were followed over some length of time.

Actually such an assumption is valid only if the plant exhibits growth of the "steady state" kind, where successive leaf and stem units can be shown to be identical. Numerous examples are known, however, where leaf and stem structures change markedly as the shoot grows. To compile a record of leaf growth under such circumstances it is necessary first to specify a certain leaf position along the shoot and then to concentrate on the development of that particular leaf.

When a leaf originates, in the form of an outgrowth from the periphery of the shoot apex, it typically takes the form of a flattened pad that shows no signs of differentiation. As its growth continues, the first two parts to be defined, in outward order, are the leaf base and the blade. The petiole, if one develops at all, arises later as an insert between the base and the blade.

The leaves of flowering plants of course have a large variety of shapes and sizes. One of the commoner variants in shape is the dissected, or compound, leaf. The blades of these leaves are cut into segments or leaflets, known to botanists as pinnae (from the Latin for "feathers"). In terms of comparative development dissected leaves are of particular interest because they so clearly illustrate how different paths of development can lead to leaves that are closely similar in appearance. The giant fronds of palm trees are the largest and most complex of all dissected leaves. At the same time the genera of the tropical arums, a family famous for its diversity of leaf forms, provide a classical example of two sharply contrasting modes of leaf dissection. Let us take up the arums first.

The mature leaves of the arum species *Zamioculcas zamiifolia* have a pinnately compound blade: four to five pairs of blade segments appear as appendages on each side of an elongated stalk, or rachis (a Latinized borrowing from the Greek for "rib"). When the terminal bud of a *Z. zamiifolia* shoot is dissected by removing successively younger leaves, a tiny dome-shaped structure, 100 micrometers in diameter, is at last re-

vealed. This is the apical meristem of the shoot; from it arise the youngest primordia, or embryonic leaves, each successively forming a small hood that arches over the shoot apex. As each new leaf primordium grows upward it completely enshrouds the apical dome and the two free edges of its hood are pressed closely together.

The growth of different tissues at different rates soon gives the leaf's future leaflets the appearance of bumplike protuberances along the leaf's two free edges. The older and slightly larger bumps appear toward the tip of the leaf; the younger and smaller ones arise progressively nearer the leaf base. Once a complement of four or five leaflet pairs has appeared, enlargement continues and the bumps soon assume the shape of mature leaflets.

This mode of leaflet initiation from the free margins of a leaf is the commonest mechanism of development among higher plants with dissected leaves, whether they are simple vascular plants such as ferns, gymnosperms such as cycads or the higher flowering plants. Through progressive degrees of elaboration it is possible for compound leaves to be twice, three times or even many times dissected. Indeed, in principle there is no limit to the degree of elaboration that can be found in a dissected leaf.

The arum that shows a sharply different mode of leaf dissection is popularly known as the split-leaf philodendron, one of the climbing genus *Monstera*. These plants are noted for the distinctive holes in the blades of their leaves and are widely grown as ornamental plants for that reason. The size and shape of the holes vary in different parts of the blade because different parts of the blade have different rates of growth. For example, the first holes to appear lie close to the edge of the blade; at maturity they are large and elliptical in outline



COCONUT PALM, probably the most numerous of all the cultivated members of the palm family, displays an abundance of compound leaves and clusters of commercially valuable nuts. The individual frond leaflets take form in a complex process of development.

because that part of the blade undergoes the greatest amount of lateral expansion. The holes that develop later are found closer to the leaf midrib. They tend to be smaller and to retain their original rounder outline because the tissue surrounding them does not undergo as much lateral growth.

It has been known for more than a century that the holes in philodendron leaves originate as localized areas of cell death. The scanning electron microscope now makes it possible to observe the process in considerable detail. The earliest sign of the impending event is the appearance of a slightly indented circular area on the surface of the leaf blade. The indentation reflects the diminished turgor (fluid content) of the affected cells. When these cells finally die, the part of the leaf composed of them simply dries up and falls away, leaving behind a perforation.

In many species of the genus *Monstera* the marginal strip of blade tissue adjacent to the most peripheral hole will not grow in width at a rate that matches the growth of the rest of the blade. As a result this thin bridge of marginal tissue usually breaks, converting what had been a perforated blade surface into a marginally lobed one. Indeed, in some species (for example *M. subpinnata, M. tenuis* and *M. dilacerata*) the leaf blade forms only a single set of broadly elliptical perforations near the blade edge. When these perforations lose their bridges, the result is a pinnately lobed leaf blade that bears a striking resemblance to the pinnate leaves of *Zamioculcas zamiifolia*. If one did not know that the dissected *Monstera* leaf had been formed by the process of cell death and the dissected *Zamioculcas* leaf by the process of lobing, the fact that each was the result of a markedly different developmental program would be impossible to guess.

Like the arums, the palms make up a large family of tropical plants that is worldwide in distribution. Not only are their fronds, or leaves, large and complex but also one species of the palm genus *Raffia* grows the largest leaves in the entire plant kingdom. Its leaflets can be 60 feet long. The palms produce a dissected leaf-blade surface in a way that is distinctly different from either of the arums' modes of leaf production. In some respects, however, the process represents a combination of the one arum's lobing and the other's cell death. Because of the greater complexity of the palm's mode of leaf production and many unresolved aspects of the process, my colleagues and I at the University of California at Berkeley have recently reinvestigated palm-leaf development in order to define the fundamentals of the process.

All palm leaves have three standard components: an elongated blade and petiole, and a leaf base that is characteristically tubular and completely encircles the stem. Among palms the leaf base plays a particularly significant role. Not only does it support the enormous and heavy photosynthetic organs of the blade and petiole; it is also a source of support for the shoot in the younger regions of the palm's stem (or trunk), where the internodes (or individual stem units) are still elongating. That is why all the bases of all palm leaves incorporate an elaborate network of fibrous tubular bundles; these provide the leaf base both with mechanical reinforcement and with flexibility.

Palm-leaf blades typically show one or the other of two major configurations that reflect differences in the distribution of growth during the course of dissection. One of the two is characterized as pinnate because of its feathery appearance. The other is characterized as palmate because it is fanlike, rather like a hand with its fingers spread. Pinnate fronds tend to have short petioles; palmate fronds have long ones.

What makes the developmental mode of palm leaves distinctive is that at first their blade surfaces are thrown into a series of pleats known as plications. A process of tissue separation along cer-





DISSECTED LEAVES of two tropical arums look much the same at maturity although they followed entirely different developmental pathways. At the left is the leaf of *Zamioculcas zamiifolia;* its five paired leaflets began as bumplike protuberances along the edge of the embryonic leaf blade (*see micrographs on opposite page*). At the right is the leaf of *Monstera subpinnata,* popularly known as the split-leaf philodendron. Its 12 leaflets originally formed a single leaf blade before the development of areas of localized cell death dissected the blade into separate leaflets (*see micrographs on page 102*). Evidence of the separate developmental pathways is not apparent at maturity.

tain of these pleats then cleaves the pleated surface into a series of leaflets. It is easy to see, in looking at a fully developed leaf of the palmate kind, that the segmented nature of the leaf originates from the pleating of the leaf blade. Indeed, as with a Japanese fan, the leaf can be closed by pressing the pleats together. Since each leaflet is *V*-shaped in cross section, it appears to have been cut out of a pleated surface.

The development of a palm leaf of the pinnate kind in general involves the same process. It is, however, subtler and more complicated, and it can best be understood by tracing the development of the leaf of the "feather" palm from inception to maturity. The lateral leaflets of the feather palm are dispersed along the leaf's rib axis; they are formed while the rib axis is elongating.

The primordium of a young feather-palm leaf arises as an outgrowth from the shoot apex. It resembles a helmet fitted over the apical dome. The part of the primordium that projects above and away from the shoot apex is the future leaf blade. The future leaf base in turn appears as a collarlike growth initiated around the periphery of the shoot apex; the shape of this early base structure foreshadows the subsequent complete encirclement of the stem by the leaf base.

The young primordium soon looks more like a hood than a helmet. Its more readily observable outer surface corresponds to the lower surface of the future leaf blade; its narrower inner surface will become the future blade's upper surface. The margins, or free edges, of the primordium represent the boundary between the two surfaces of the future blade.

Where in the lobe-growing arum *Zamioculcas* the initiation of leaflets involves lobing along this leaf-blade edge, the feather palm's future leaflets first appear as a series of ripples or ridges within the surface of the blade some distance from the edge of the blade. The folds are not only particularly evident on the readily observable lower blade surface-to-be but also have their counterparts on the upper surface. A section made at a right angle to one of the blade margins cuts across these ridges and reveals that the lower and upper ridges actually make up a single set of pleats running the entire length of the blade in a compressed zigzag.

What makes these pleats unique in leaf development is their spatial restriction. For example, they never extend to the blade margins, with the result that the growing leaf carries a narrow unfolded strip of tissue around the entire periphery of the blade. Similarly, on the back side of the leaf the pleats extend only a short distance toward the axis of the thickened leaf rib.

As the leaf continues to grow, new pleats appear. Most of them are added toward the base of the leaf blade but a few may appear toward the tip of the blade as well. The total number of pleats typically corresponds to the number of leaflets found in the mature leaf, and so the development of additional pleats is terminated when that number has been attained. At this stage the pleats remain tightly pressed together, like the folds of a camera bellows.



EMBRYONIC LEAF of *Z. zamiifolia* is seen at an early stage in the photomicrograph at the left and at a slightly later stage in the photomicrograph at the right. The specimens are 1.7 and three millimeters long. Bumps on the free margins of the younger leaf are more recognizable as emerging leaflets on the older one; they are starting to approximate their mature shape.

The next and most intricate steps in the development of the feather-palm leaf take place after the pleats have undergone further growth in depth. The steps are two: separation of the pleats into individual leaflets and disposal of the peripheral leaf-blade tissue that still binds the leaflets together at their tips. In the first step the adjacent pleats separate from each other. Although the cellular features of the process are not yet known in detail, it is assumed that separation is effected by a breakdown of the intercellular cementing substance along the line of the separation. It is certain that selective cell death is not involved: there is no evidence of cell damage along the separation.

Before proceeding to the second step in leaflet formation some general features of palm-leaf development in the first step should be mentioned. One of them is that in the majority of pinnate-leaved palm species the separation process takes place along the ridges of the outer, or lower, leaf-surface-to-be. The process gives rise to individual *V*-shaped leaflets. Observation of the process by scanning electron microscopy shows that an indentation develops along the top of the ridge. As tissue separation takes place the ridge indentation is transformed into a distinct slit.

Returning to the second step, even though the leaflets have been separated from one another over most of their length, their tips remain attached to the strip of leaf-edge tissue. The condition does not prevail for long, although even after the breakup of the tissue begins, strips of the nearly separated margin may still remain attached to the tips of individual leaflets. Because of their resemblance to parts of a horse's bridle, these dangling strips are called the "reins" of the leaf.

The mode of origin of palm-leaf pleating has been a subject of contro-



LEAF PERFORATIONS that will eventually dissect the blade of a *Monstera* leaf into individual leaflets can be seen in this scanning electron micrograph, which magnifies a developing leaf 35 times. In the younger part of the leaf, at the right, the perforation-to-be appears as an indentation on the surface of the leaf blade. In the older part, at the left, some of the dying tissue has broken away from the surrounding edge of still healthily turgid leaf-blade cells.



LATER STAGE OF LEAF PERFORATION is shown in a scanning electron micrograph of a *Monstera* leaf blade. The dead tissue has fallen away. As the leaf blade continues to grow, the tissue at the bottom, between the lower edge of the perforation and the blade margin, will tear apart, turning the two sides of the perforation into the lower edge of one leaflet and the upper edge of the adjacent one. This *Monstera* specimen, 17 millimeters long, is magnified 35 times.

versy in botanical circles for a century and a half. Each of two separate developmental hypotheses, representing radically different morphogenetic alternatives, has had until recently its own rank of supporters. The simplest and most direct of the two hypotheses proposes that pleating results from an initial folding produced by differential growth within the young leaf blade.

The second hypothesis proposes that pleating begins with an initial process of tissue separation, which is then followed by differential growth. According to this hypothesis, alternating clefts develop along both the upper and the lower surfaces of the blade because progressive cell separation begins on the blade surface and forms a furrow that extends to the interior of the leaf. As the cells separate from one another this interior penetration deepens until the resulting zigzag outline of the blade is difficult to distinguish from an outline produced by folding, particularly after the folds have become compressed.

What makes the second hypothesis so much more complex than the first is that it calls for some remarkable feats of cell differentiation. For example, consider the layer of cells on the surface of the leaf blade: the leaf epidermis. According to the hypothesis, the development of furrows would divide the leaf epidermis into isolated surface patches. If, as is observed, the sides of the furrows are also covered by a layer of epidermal cells, then some of the cells must have regenerated from cells deep in the interior of the leaf. These interior cells, however, would not normally have been destined to become epidermal cells. Is such a regeneration possible? If it is, it would represent a mechanism unique among higher plants.

The question was sufficiently intriguing to enlist Nancy and Ronald Dengler of the University of Toronto as additional allies in my investigation. Although what I shall hereafter refer to as the "tissue splitting" hypothesis lacked unequivocal evidence in its favor, there remained several reasons such a mechanism might in fact be operating in palms. The first was that some kind of tissue separation was clearly involved in the cleavage of palm-leaf pleats into individual leaflets. This being so, could not the same process be responsible for the origin of the pleats in the first place? William of Occam would surely have approved of a single developmental program rather than two different ones.

A second datum in favor of tissue splitting was an observation made as young palm leaves were being examined by scanning electron microscopy: the initial pleat furrows looked very much like narrow slits cut into the leaf surface. This fact, together with the observation that in some palm species the pleats ap-

pear to be block-shaped and are separated by narrow slits, further biased us in favor of the tissue-splitting hypothesis.

A survey of the scientific literature showed that one reason the two hypotheses had remained unresolved for so long was that earlier investigators, lacking more powerful instrumentation, had necessarily accepted as definitive data the superficial appearances of pleating shapes. They took these indications as proof of how the process occurred rather than relying on more fundamental measurements that were independent of surface appearances.

We set as our goal the study of the cellular and anatomical bases of pleating origins in the leaves of several different palm species, utilizing scanning electron microscopy in addition to light microscopy in order to pay particular attention to the three-dimensional aspects of leaf-blade development. We also studied thin sections of plant tissue (1.3 micrometers thick), prepared by the same techniques applied in transmission electron microscopy. (In transmission electron microscopy, unlike scanning electron microscopy, the electrons must pass through the specimen.) The use of these thin sections gave us light-microscope images of superior resolution, enabling us to eliminate many of the light-microscope artifacts, or misleading false elements, that had plagued earlier investigators.

Now, according to both the tissue-splitting and the differential-growth hypotheses on the origin of leaf pleats, the furrows that separate the pleat ridges must deepen as growth continues. In the tissue-splitting model the furrows themselves would deepen. In the differential-growth model the increase in depth would be caused by growth of the ridges upward from the base of the furrow. Thus it should be possible to determine unequivocally which of the two models was operative by simply counting the number of cells that remained in contact in the lower and upper pleat ridges, either below or above the respective furrow lines.

For example, we predicted that if tissue splitting was the process in action, then the number of cells remaining in contact opposite the base of a furrow should decrease over time because cells were being pulled apart by the same forces of cell separation that were deepening the furrow. We also predicted that if the increased depth of the furrow were the result of ridge upgrowth, in accord with the differential-growth model, then the number of cells in contact along the furrow line should either increase (if the growth of the adjacent ridges involved both cell division and cell enlargement) or remain the same (if the growth involved cell enlargement alone).

Our counts of cells proved that the

Friends of Libraries U.S.A.

We're a new national organization dedicated to encouraging and assisting the formation and development of Friends of Libraries groups throughout the country.

Won't you join us?

Name

Address

City State Zip

Library

Address

City State Zip

For an application form, write:
Friends, ALA
50 E. Huron
Chicago, Illinois 60611

LYNCHBURG HARDWARE & GENERAL STORE®
47 Main St., Lynchburg, TN 37352



JACK DANIEL SPORT SHIRT

This white, top quality shirt is made of 50% cotton and 50% polyester. It sports a flap pocket embroidered with the Jack Daniel emblem in black and gold. Perfect for the golf course or just sporting around. Order S, M, L or XL. **$20.00** delivered.

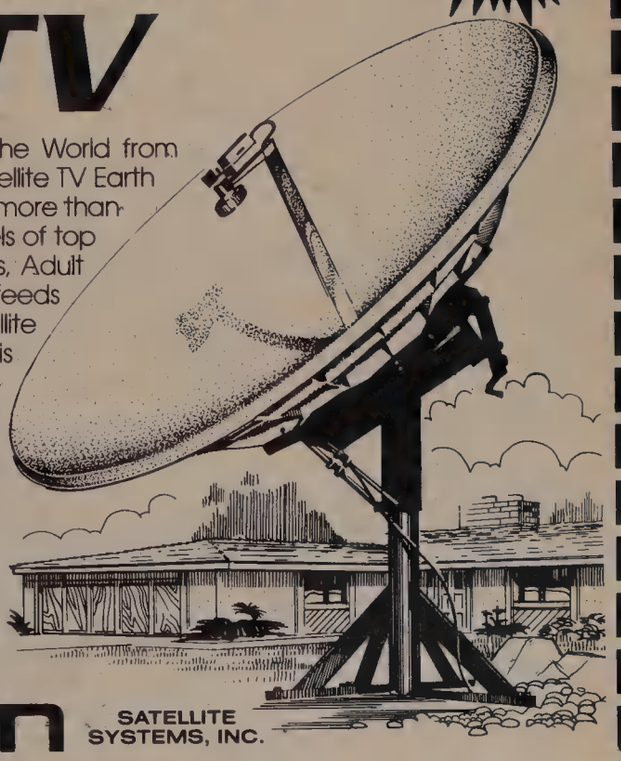
Send check, money order or use American Express, Diners Club, Visa or MasterCard, including all numbers and signature. (Add 6¾% sales tax for TN delivery.) For a free catalog, write to Eddie Swing at the above address. Telephone: 615-759-7184

Now, developed exclusively for the home-

Satellite TV

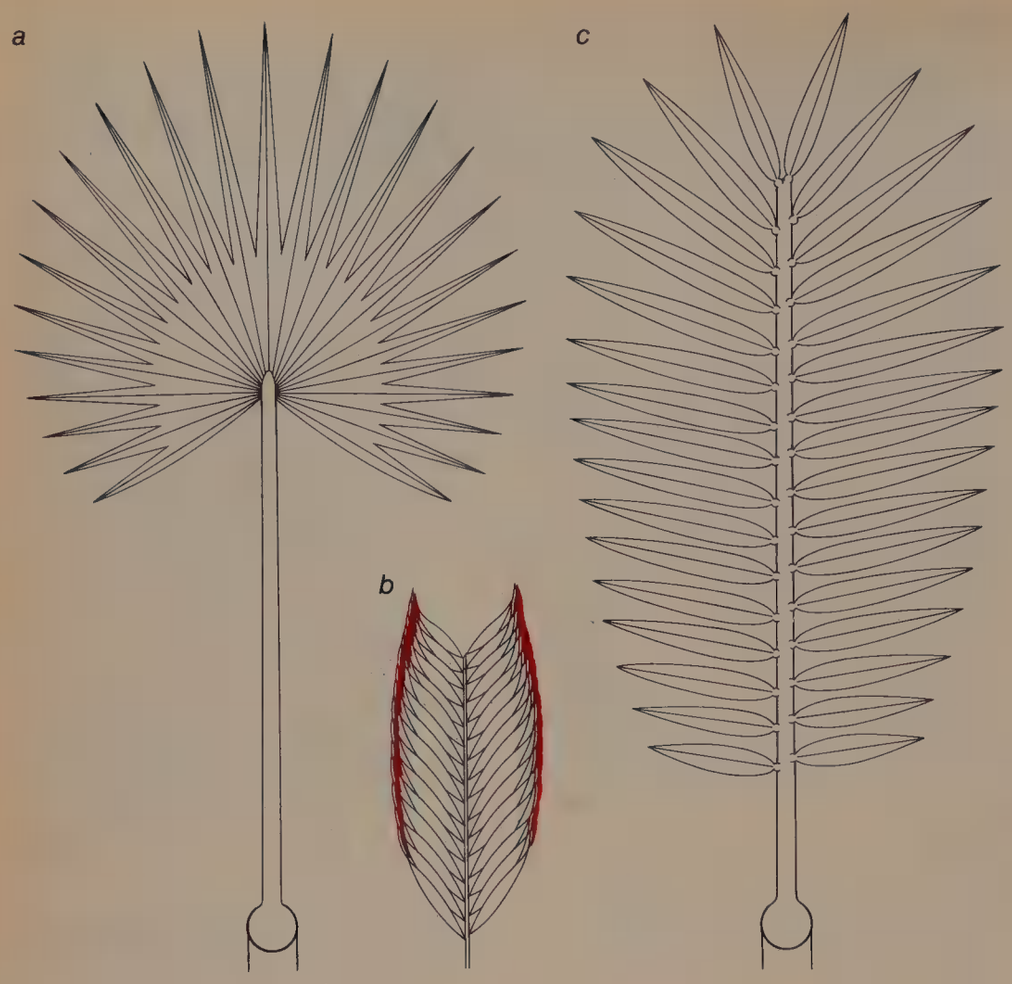
WE SHIP ANYWHERE!

Now you can tune in the World from home with your own Satellite TV Earth Station System. Receive more than 80 earth-bound channels of top Movies, live sports, News, Adult Programming,Cable TV feeds and more – all via Satellite 24 hrs./day. Installation is simple with our video-taped instructions or if you prefer – we can often arrange installation for you! Prices for systems begin at a remarkable **$2,295** and all come with a full 1 year parts & labor warranty. NO GIMMICKS. Call or write for full information.

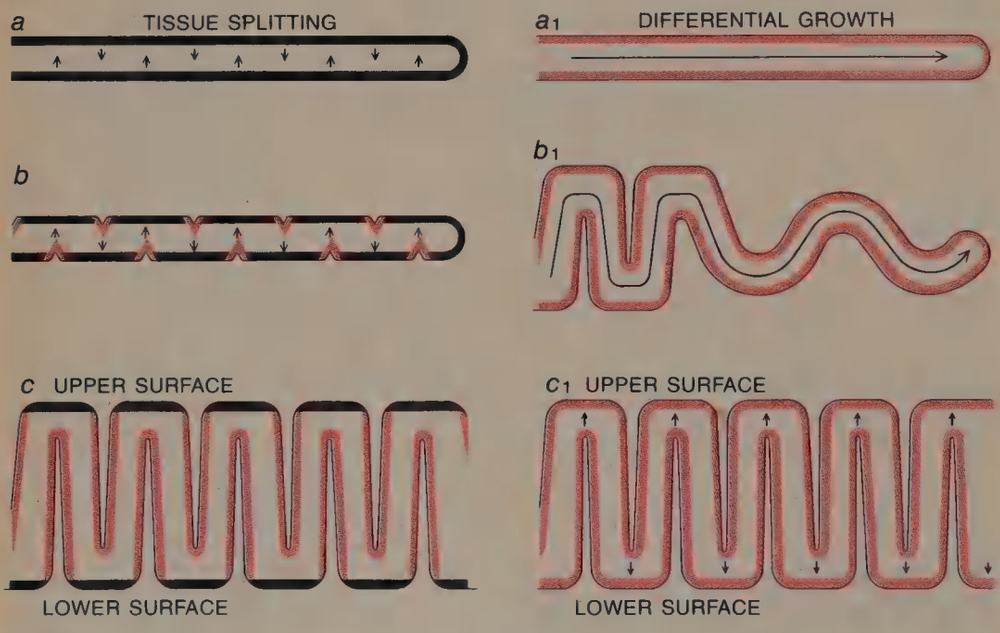


ORION SATELLITE SYSTEMS, INC.

3320 SIRIUS ST. • LAS VEGAS, NV 89102 • (702) 362-9992



MATURE PALM LEAVES generally show either a fanlike form (*a*), with a long petiole separating the leaf blade from the leaf base, or a featherlike form (*c*), with a short petiole and leaflets on each side of a long central rib. Before either array of leaflets can unfold (*b*) it must shed its "reins" (*color*): the thin strips of leaf-blade margin connecting the tips of the leaflets.



PALM-LEAF PLEATING had been thought to be due to one of two mutually exclusive patterns of development. The first pattern, shown in three stages at the left in this diagram, would initiate (*a*) as "tissue splitting" (cell separation) on the upper and lower surfaces of the leaf blade, a process that would divide the epidermal layer of the blade (*gray*) into unconnected units (*b*). As leaf-blade growth continued (*c*) the zones of mesophyll tissue connecting the upper and lower surfaces would expand substantially (*color*) and somehow regenerate as epidermal cells. The alternative pattern of development, at the right, visualizes pleating as the result of "differential growth" (a_1, b_1), whereby the pleats would first arise from crowding during growth and thereafter be extended by upward and downward processes of growth within the leaf blade, maintaining an unbroken layer of epidermal tissue (*color*) throughout the process (c_1). The author and his colleagues, studying scanning electron micrographs and thin sections of different palm leaves, concluded that the tissue-splitting hypothesis could not be supported because no cell separation was visible, whereas cell growth and enlargement could be seen.

deepening of the furrow lines was the result not of cell separation but of ridge upgrowth. Although the process was not identical for the lower and upper pleat ridges, no decrease in cell numbers (indicative of splitting) was evident in either. In the upper ridges cell numbers initially increase; in the lower ridges the numbers remain about the same. Further observations gave additional support to the differential-growth hypothesis. For example, we found that the pleats had a continuous epidermal layer over their folded surfaces at all stages of development. By the same token there was no evidence at any stage of a splitting and redifferentiation of the epidermal layer.

Our studies included both palmate and pinnate palm species and gave virtually identical results regardless of the species' blade morphology. Therefore we suggest that our findings are probably valid for members of the palm family in general. One other issue, however, still remained unresolved. Why did the primordial folds in the leaves of some palm species assume a shape superficially suggestive of tissue splitting? Having been misled ourselves, we were anxious to find out, once the pleats were initiated, what forces might be influencing their shape.

It had been suggested in the 19th century that space limitations within the terminal bud were responsible for the initial pleating of palm-leaf primordia. The proposal was almost certainly inspired by the sight of tightly folded young leaves in the crown of palms. Could this proposed relation between pleating and space restrictions, obvious in the later stages of leaf development, apply to the stage where the leaf primordia were virtually microscopic in size and crowding seemed unlikely? Circumstantial evidence suggests not. For example, if pleating were simply a response to space limitations, one might logically expect the entire leaf-blade surface to be folded. Yet the margins of the leaf blade are never folded. What is more, when one analyzes the relation between the space within the sheath of the next-older leaf and the shape and distribution of pleats on the blade of the next-younger leaf, one finds no correlation between the space available for expansion and the orientation and extent of the pleats.

This is not to say that space limitations in the bud do not have an effect on the shape of the pleats. For example, in palms that have the bud leaves tightly packed together the outer leaf surface appears flat and the pleat furrows resemble slits cut into the leaf blade. In palms that have less tightly compacted bud leaves, however, the pleats bulge quite conspicuously, giving the impression that they arise as a result of blade

folding rather than of tissue splitting.

An acid test of the possible role of crowding in leaf-blade pleating would be the growth of isolated leaf primordia in a nutrient medium to allow observation of their development in an uncrowded environment. Pilot experiments have established the possibility of rearing such primordia in vitro from an early stage, when they are less than a millimeter in length. We have not yet, however, undertaken a rigorous analysis of their development to see how it may compare with the situation in vivo.

Our studies of palm-leaf development were inspired by the suggestion that the leaf blade's dissection into leaflets involved a process unique in the plant kingdom: tissue splitting. The investigation established, however, that whereas a process of tissue separation is indeed sometimes responsible for cleavage of the leaf blade's initial folds into distinct segments, the development of a pleated leaf blade involves a process no more complicated than differential growth.

In essence, then, the morphogenetic process responsible for palm-leaf dissection is really much the same process responsible for the more typical development of leaflets in the arum *Zamioculcas*. The two processes differ only in the site where the differential growth takes place. In the arum this site is at the free edge of the leaf blade, whereas in the palm the growth takes place within the surface of the blade at some distance from the blade edge. Indeed, one can argue that all the complexity characteristic of palm-leaf development, from pleating and tissue separation along selected pleats down to the final abscission of the reins, is a consequence of leaflet initiation within the blade rather than at its edge. Viewed in this light the many idiosyncrasies of palm-leaf development seem far less aberrant.

In the long run, of course, the interests of botany will be poorly served by any mere cataloguing of developmental pathways in plants. Instead the botanist's ultimate goal must be the coupling of such developmental information with the rich data from the molecular levels of biological organization. For example, most molecular biologists assume that the systems of control they have diligently traced in plants are to be found in all the higher plants. This may not be true. Perhaps the different kinds of development described here are indicative of greater differences among molecular control systems than seem apparent now. It may well be that future studies of morphogenesis will play a significant role in investigators' thinking about the mechanisms of plant growth and development only if they effectively integrate information available from the other fields of biological research.



PLEATING FOLDS in the maturing leaf blade of a pinnate palm of the genus *Chamaedorea* are seen in this scanning electron micrograph. The specimen, three millimeters long, is magnified 45 times. The inner ridges at the right, like those partly visible just to the left, will become the midribs of each of the leaflets-to-be. Some outer leaflet ridges, visible only to the left, show developing central axial depressions. Tissue cleavage will continue along these depressions until each of the leaflets is separated from the next (*see the further advanced specimen below*).



LEFT OUTER RIDGES of the same pinnate leaf blade are seen in a scanning electron micrograph at the same magnification. The tissue splitting of the outer ridges has progressed and the relation between the leaflet halves-to-be and their respective midribs is more readily evident.

Particles with Naked Beauty

The fifth quark, embodying the "flavor" known as beauty, has now been seen in combination with an antiquark not of its own flavor. The beauty of the new composite particles is accordingly exposed

by Nariman B. Mistry, Ronald A. Poling and Edward H. Thorndike

The ultimate building blocks of matter are currently thought to be the small set of indivisible particles called leptons and quarks. The world we normally experience is composed almost entirely of one kind of lepton, the electron, and two kinds of quark, arbitrarily labeled "up" and "down" (or simply u and d). The fractionally charged quarks, which are distinguished from one another by the quantum properties known as flavors, are bound together in different combinations to constitute the particles of the atomic nucleus: the proton (which has the composite structure represented by the letters uud) and the neutron (udd).

A complete description of all the known subatomic particles, including those observed only fleetingly in high-energy particle accelerators, requires more of these fundamental constituents. Six kinds of lepton are recognized, falling naturally into three doublets, or pairs. Until recently four kinds of quark were also well established, forming two similar doublets. In addition to the u and d quarks of ordinary matter a second pair of quarks is needed to embody the special flavors termed strangeness (s) and charm (c). This article summarizes the experimental evidence for a fifth kind of quark: the bottom, or b, quark, which embodies the flavor referred to as beauty.

The first indication of the existence of quarks endowed with beauty came six years ago as the result of experiments in which a beam of high-energy protons was directed against a stationary target. Among the products of the bombardment were a number of massive two-quark particles that appeared to consist of a new quark, the b quark, bound to its own antiparticle, the $\bar{b}$ quark. In such a particle the opposite flavors cancel and the particle is said to be flavorless; in other words, the observed particles were judged to have hidden, or concealed, beauty.

For the past few years our group at Cornell University has taken another approach to the search for particles with beauty. In our experiments oppositely directed beams of high-energy electrons and positrons are brought into collision; the colliding particles of matter and antimatter annihilate each other in a flash of electromagnetic radiation from which new particles can materialize. With this method we have succeeded in creating particles in which the b quark is apparently bound to an antiquark of a different flavor. In these particles the beauty is no longer hidden; indeed, it is said to be naked. The observation of particles with naked beauty is expected to yield new information on the relations among the fundamental particles. It is also expected to focus renewed attention on the search for yet another quark to complete the third quark doublet. The predicted but so far unobserved sixth quark is labeled the top, or t, quark, and its flavor is referred to as truth. Before telling the story of the discovery of the beauty quark we shall review in somewhat greater detail the evolution of the lepton/quark description of matter.

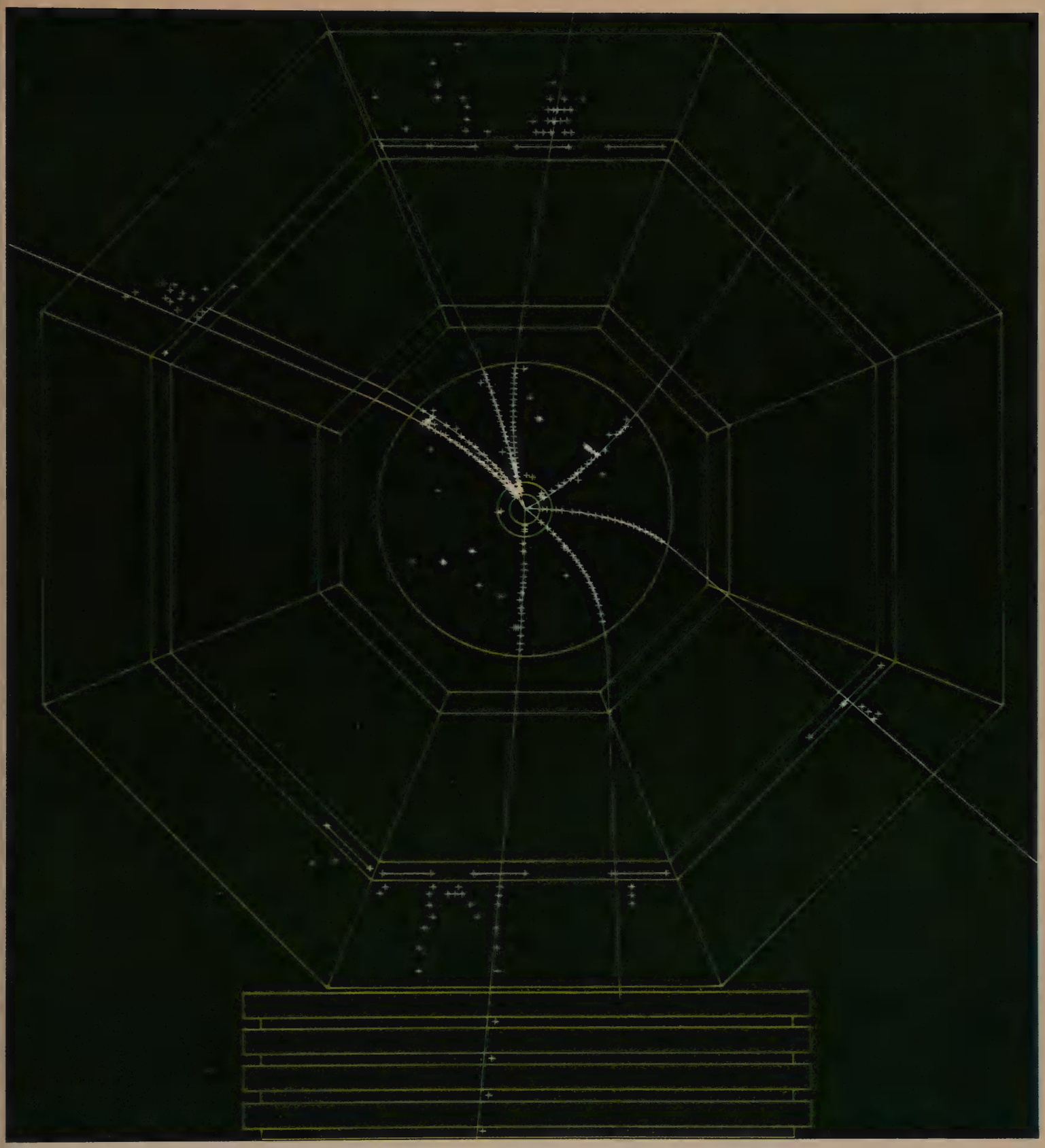
The three lepton doublets consist of the electron (e^-) and its uncharged partner, the electron neutrino (ν_e), the muon (μ^-) and its neutrino (ν_μ) and the tau (τ^-) and its neutrino (ν_τ). For each lepton there is an antilepton, and so the full membership of the class also includes the positron (e^+) and its antineutrino ($\bar{\nu}_e$), the positive muon (μ^+) and its antineutrino ($\bar{\nu}_\mu$) and the positive tau (τ^+) and its antineutrino ($\bar{\nu}_\tau$). The leptons are very small, no more than 10^{-15} centimeter in diameter, and are probably pointlike. Although the charged leptons behave alike, their masses are quite different. The muon is 206 times as massive as the electron, and the tau is almost 3,500 times as massive as the electron. The neutrinos are very light and may actually be massless.

Each of the three quark doublets (u and d, s and c, b and t) is made up of one quark with charge 2/3 and another with charge $-1/3$. Like the leptons, the quarks seem very small and are usually thought of as point objects. Similarly, for each quark there is an antiquark, with opposite charge and opposite flavor. For example, since the b quark has charge $-1/3$ and -1 unit of beauty, its antiquark, the $\bar{b}$, has charge $+2/3$ and $+1$ unit of beauty.

With 24 building blocks (leptons, antileptons, quarks and antiquarks) it is quite remarkable that under most circumstances matter can be accounted for with only one lepton doublet (e^-, ν_e) and one quark doublet (u, d). In addition to the three fundamental particles already mentioned as the main constituents of ordinary matter the fourth member of these first two doublets, the electron neutrino, is emitted by matter as a by-product of natural radioactivity.

Quarks are bound in other combinations to make up many particles besides protons and neutrons. Two possible combinations exist: groups of three quarks, called baryons, and quark-antiquark pairs, called mesons. Together the two classes of particles built from quarks are called hadrons. Hadrons, although they are small, are not pointlike. Their diameter of about 10^{-13} centimeter reflects their composite structure. Because of the strength of the force that binds quarks together, quarks are not detected in isolation. Physicists learn about quarks by studying hadrons. More than 100 different hadrons are known; most of them, however, are seen only in high-energy collisions produced by particle accelerators.

Of the four forces of nature three—the strong force, the electromagnetic force and the weak force—play important roles in the interactions of leptons and quarks. The fourth force, gravity, which is important in the macroscopic world, is negligible on the scale of quarks and leptons. In the modern view of the forces two particles interact by exchanging entities called gauge bosons. The exchanged objects behave somewhat like particles, but they do not have well-de-



BEAUTY-FLAVORED MESONS (that is, particles made of two quarks, only one of which is endowed with the property called beauty) are responsible for the pattern of particle tracks represented in white in this computer-generated display. The origin of the tracks was a head-on collision between an electron and a positron. The colliding particles of matter and antimatter had been accelerated in opposite directions within the circular vacuum chamber of the Cornell Electron Storage Ring (CESR). The collision took place inside the CLEO detector, one of two large particle-identification systems set up at the interaction points where the counterrotating beams cross. (A plan view of the accelerator and a three-dimensional view of the CLEO detector appear on page 111.) The mutual annihilation of the electron and the positron resulted in a flash of electromagnetic radiation, from which there promptly materialized a massive unstable meson called the upsilon (Υ'''), consisting of a beauty quark (b) loosely bound to its own antiquark ($\bar{b}$). In the subsequent decay of the Υ''' two beauty-flavored mesons, designated B and $\bar{B}$, were formed, and they in turn decayed into an assortment of other particles. The passage of the electrically charged decay products through various parts of the detector was recorded as a pattern of "hits" (*white crosses*). By fitting tracks to the hits an automatic pattern-recognizing computer program determined the direction and momentum of each particle and in some cases identified the type of particle. For example, the track at the six-o'clock position was attributed to an extremely penetrating muon because of the string of hits registered in a special set of detection chambers arrayed around the outside of the detector. The track at 12 o'clock was identified as a probable electron on the basis of information recorded about the rate at which it deposited energy in an inner set of chambers, together with the telltale "shower" pattern it left in the electromagnetic-shower chamber at the top of the detector. A complete analysis of such collision events relies on data from all the detector's components, including some not shown here.

fined masses and they exist very briefly.

The strong force acts only among quarks; it is transmitted by the gauge bosons called gluons, which are so named because they are the "glue" that binds quarks together into hadrons. The electromagnetic force acts between any pair of charged particles; it is transmitted by the photon, which is the massless quantum of electromagnetic radiation. The weak force affects both quarks and leptons; it is transmitted by the gauge bosons designated W^+, W^- and Z^0 (also known as intermediate vector bosons). Since the strong force is the strongest of the three forces, it plays the dominant role in quark and hadron processes, unless it is inhibited for some reason. In contrast, the weakness of the weak force, in the range of energies currently being examined, ensures that it is significant only when the strong and electromagnetic forces are suppressed. As the available energy increases, the weak force becomes stronger, until it eventually becomes comparable in strength to the other forces.

The gauge bosons associated with the respective forces act as the agents in all processes in which those forces are involved. The strong bonding of a d and a $\bar{u}$ quark to form a negative pion, or π^- meson, for example, is attributable to the exchange of gluons between the two quarks. Similarly, the electromagnetic force between two electrons is mediated by the exchange of photons. Another kind of electromagnetic interaction is observed when an electron and a positron collide and annihilate each other to form a photon, which promptly materializes into a positive muon and a negative muon. The weak interaction is responsible for processes such as the decay of a tau lepton, in which the tau is transformed into a tau neutrino by emitting a W^- boson. The W^- in turn decays into a negative muon and a muon antineutrino.

The heavier quarks and leptons are unstable and decay into lighter quarks and leptons. The quark decays entail a change in flavor. Gluons and photons are insensitive to flavor and cannot cause such changes; hence the quark and lepton decays are classified as weak interactions. In the decay of the s quark, for example, the s quark becomes a u quark when it emits a W^-. The W^- in turn can materialize into several possible pairs of particles, such as an electron and an electron antineutrino. One might also expect to observe the decay of an s quark into a d quark with the s quark emitting a Z^0 (the neutral counterpart of the W^+ and W^-), which could then decay into an electron and a positron. Careful searches have shown that such processes involving the Z^0 do not occur.

The need to incorporate this experimental fact into the theory led to the hypothesis that the s quark is the charge $-1/3$ member of a doublet with a charge $2/3$ quark of a new flavor, labeled charm. With two quark doublets the interplay of the quarks is such that the probability of the decay of an s quark into a nonstrange quark by the emission of a Z^0 vanishes. The existence of the c quark was predicted on this basis almost 20 years ago, 10 years before the dramatic discovery of psi (ψ) particles, which are mesons composed of a c quark and a $\bar{c}$ antiquark [see "Electron-Positron Annihilation and the New Particles," by Sidney D. Drell; SCIENTIFIC AMERICAN, June, 1975]. This mechanism for suppressing certain strangeness-changing processes can be generalized to other flavors. In particular, if the b quark shares a doublet with a t quark, beauty-changing processes involving a Z^0 are forbidden. If there is no t quark, however, such processes must happen.

After the charmed quark was discovered in 1974 the lepton/quark description of matter seemed to be on a sound footing. The two lepton doublets and the two quark doublets recognized by that time accounted nicely for all the known particles. There was no compelling reason to expect any other leptons or quarks. The discovery of the tau lepton in 1975, however, suggested that there might be more quarks too [see "Heavy Leptons," by Martin L. Perl and William T. Kirk; SCIENTIFIC AMERICAN, March, 1978]. Evidence for a fifth quark was not long in coming. Indeed, there are many parallels between the discovery of the c quark and that of the b.

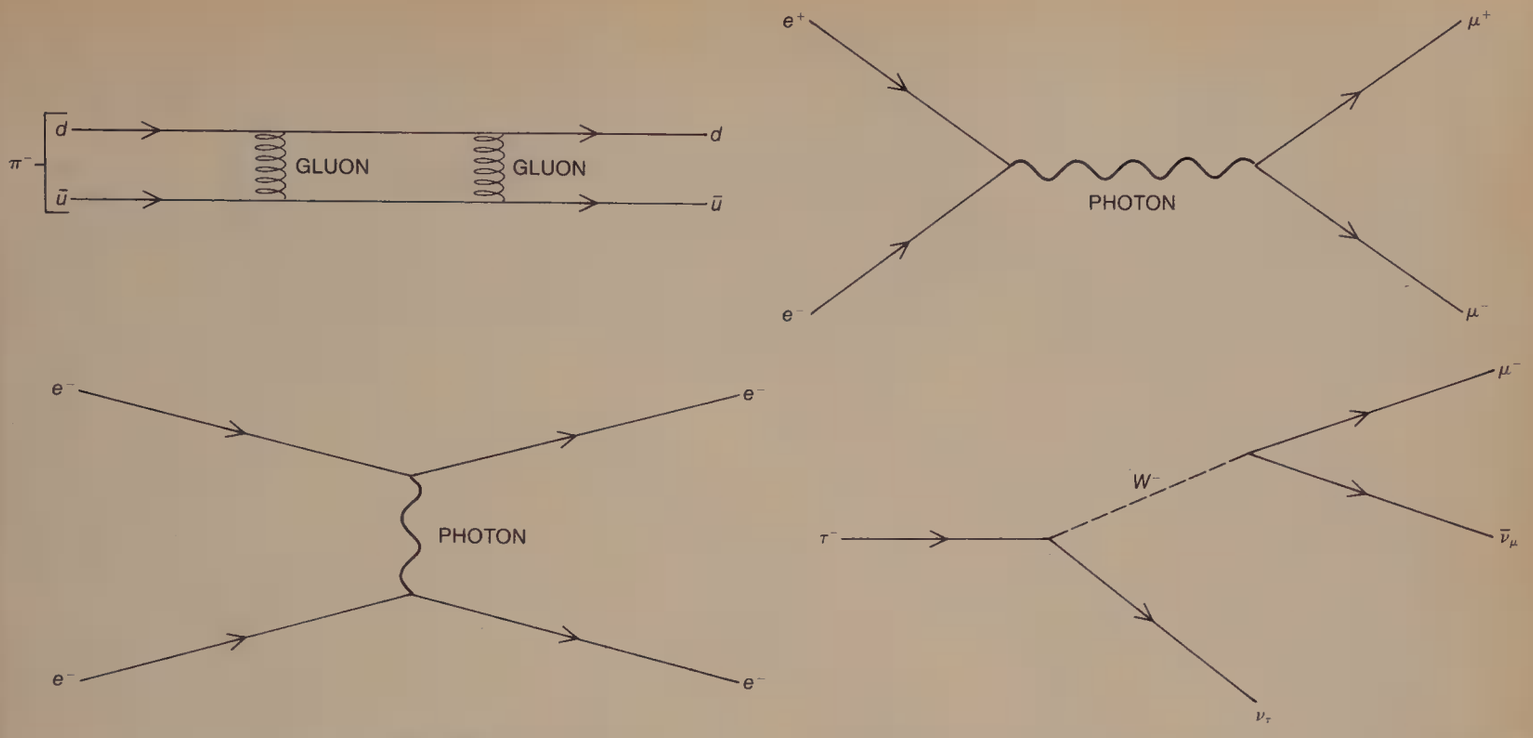
The first piece of evidence resulted from a 1977 experiment by a group of physicists at the Fermi National Accelerator Laboratory. The experimenters, led by Leon M. Lederman, then of Columbia University, bombarded a copper target with a beam of protons at an energy of 400 billion electron volts (GeV) and searched for muons produced in the collisions. When they observed a pair of oppositely charged muons, they considered the possibility that a new particle had been created and had subsequently decayed into the $\mu^+\mu^-$ pair. By measuring the momentum and the direction of each of the muons the experimenters were able to determine the mass of the new particle. The mass distribution that was obtained showed a broad peak at 10 GeV, indicating the presence of the new particle, which was named the upsilon (Y). Lederman and his colleagues interpreted the shape of their peak as showing that they were in fact seeing two or more narrow peaks (probably three), smeared together by the limited experimental resolution [see "The Upsilon Particle," by Leon M. Lederman; SCIENTIFIC AMERICAN, October, 1978].

The decay of the new particle into a $\mu^+\mu^-$ pair is an electromagnetic process. If the upsilon were to decay by the strong force acting at full strength, the electromagnetic decays would be insignificant in comparison and the $\mu^+\mu^-$ decay would not have been detected. Its

CHARGE	LEPTON DOUBLETS			ANTILEPTON DOUBLETS	CHARGE
0	ELECTRON NEUTRINO	ν_e	$\bar{\nu}_e$	ELECTRON ANTINEUTRINO	0
−1	ELECTRON	e^-	e^+	POSITRON	+1
0	MUON NEUTRINO	ν_μ	$\bar{\nu}_\mu$	MUON ANTINEUTRINO	0
−1	NEGATIVE MUON	μ^-	μ^+	POSITIVE MUON	+1
0	TAU NEUTRINO	ν_τ	$\bar{\nu}_\tau$	TAU ANTINEUTRINO	0
−1	NEGATIVE TAU	τ^-	τ^+	POSITIVE TAU	+1

CHARGE	QUARK DOUBLETS			ANTIQUARK DOUBLETS	CHARGE
+2/3	UP	u	$\bar{u}$	ANTIUP	−2/3
−1/3	DOWN	d	$\bar{d}$	ANTIDOWN	+1/3
+2/3	CHARMED	c	$\bar{c}$	ANTICHARMED	−2/3
−1/3	STRANGE	s	$\bar{s}$	ANTISTRANGE	+1/3
+2/3	TOP (TRUTH)	t	$\bar{t}$	ANTITOP (ANTITRUTH)	−2/3
−1/3	BOTTOM (BEAUTY)	b	$\bar{b}$	ANTIBOTTOM (ANTIBEAUTY)	+1/3

LEPTONS AND QUARKS, the two classes of indivisible, pointlike particles currently assumed to be the ultimate building blocks of matter, fall naturally into pairs called doublets. The leptons, which have integer charges (−1, +1 or 0), are observed routinely in the laboratory. The fractionally charged quarks are not observed in isolation; they apparently exist only as the constituents of the composite particles known as hadrons. For each lepton or quark there is a corresponding antilepton or antiquark with the same mass but with opposite charge and flavor.



THREE FORCES important in the interaction of subatomic particles operate by the exchange of intermediary particles called gauge bosons. In these schematic diagrams showing how such forces work time goes from left to right. The gluon, the gauge boson of the strong force, is responsible for binding the quarks together into hadrons; in this case the d quark and the $\bar{u}$ quark that constitute a negative pion, or π^- meson, are tightly bound by the periodic exchange of a gluon (*diagram at upper left*). The photon, the massless quantum of electromagnetic radiation, serves as the carrier of the electromagnetic force; for example, when two electrons approach each other, they interact electromagnetically by exchanging a photon (*diagram at lower left*). In another kind of electromagnetic interaction an electron and a positron collide and annihilate each other to form a photon, which can then materialize into new particles, such as a muon-antimuon pair (*diagram at upper right*). Leptons and quarks decay through the weak interaction. Thus when a negative tau lepton decays, it is transformed into a tau neutrino by the emission of the intermediate vector boson labeled W^-; the W^-, one of three such carriers of the weak force, decays into one of several possible combinations of particles, such as a negative muon and a muon antineutrino (*diagram at lower right*).

observation suggested that the strong decay of the upsilon is somewhat inhibited. It was recognized that if the upsilon were a meson composed of a new quark and its antiquark ($b\bar{b}$), such a suppression could be readily explained. With enough energy the b and $\bar{b}$ quarks in the meson could separate, creating a new quark-antiquark pair and yielding "beauty-flavored" mesons, for example $b\bar{u}$ and $\bar{b}u$.

The latter decay is a strong-interaction process that would not be suppressed if it were energetically allowed. The observed suppression suggests that there is too little energy available for this process; in other words, the mass of the upsilon is less than the sum of the masses of two beauty-flavored mesons. The only alternative strong-decay mechanism is for the b and $\bar{b}$ quarks to annihilate each other and produce hadrons. This type of annihilation, although it is mediated by the strong interaction, is known to be suppressed. Indeed, it is because of the suppression that the electromagnetic decay is important. In the electromagnetic process the b and $\bar{b}$ quarks annihilate each other to form a photon, which then materializes into a $\mu^+\mu^-$ pair.

The suppression of the strong decay is expected to result in a lifetime for the upsilon that is considerably longer than the 10^{-23}-second lifetime of a typical strongly decaying hadron. Time intervals on this order are still too brief to be measured directly, but they can be determined from a relation between the lifetime of a particle and the precision to which its mass can be measured. If a particle lives forever, its mass can be determined with perfect accuracy. If its lifetime is very brief, measurements of its mass will give a distribution with a width related to the value of the lifetime; for a normal strongly decaying hadron the width of the mass distribution would be equivalent to an energy range of about .15 GeV. Lederman's experiment did not have sufficient resolution to measure the width of the upsilon's mass-distribution peaks; this determination awaited the arrival of a different technique.

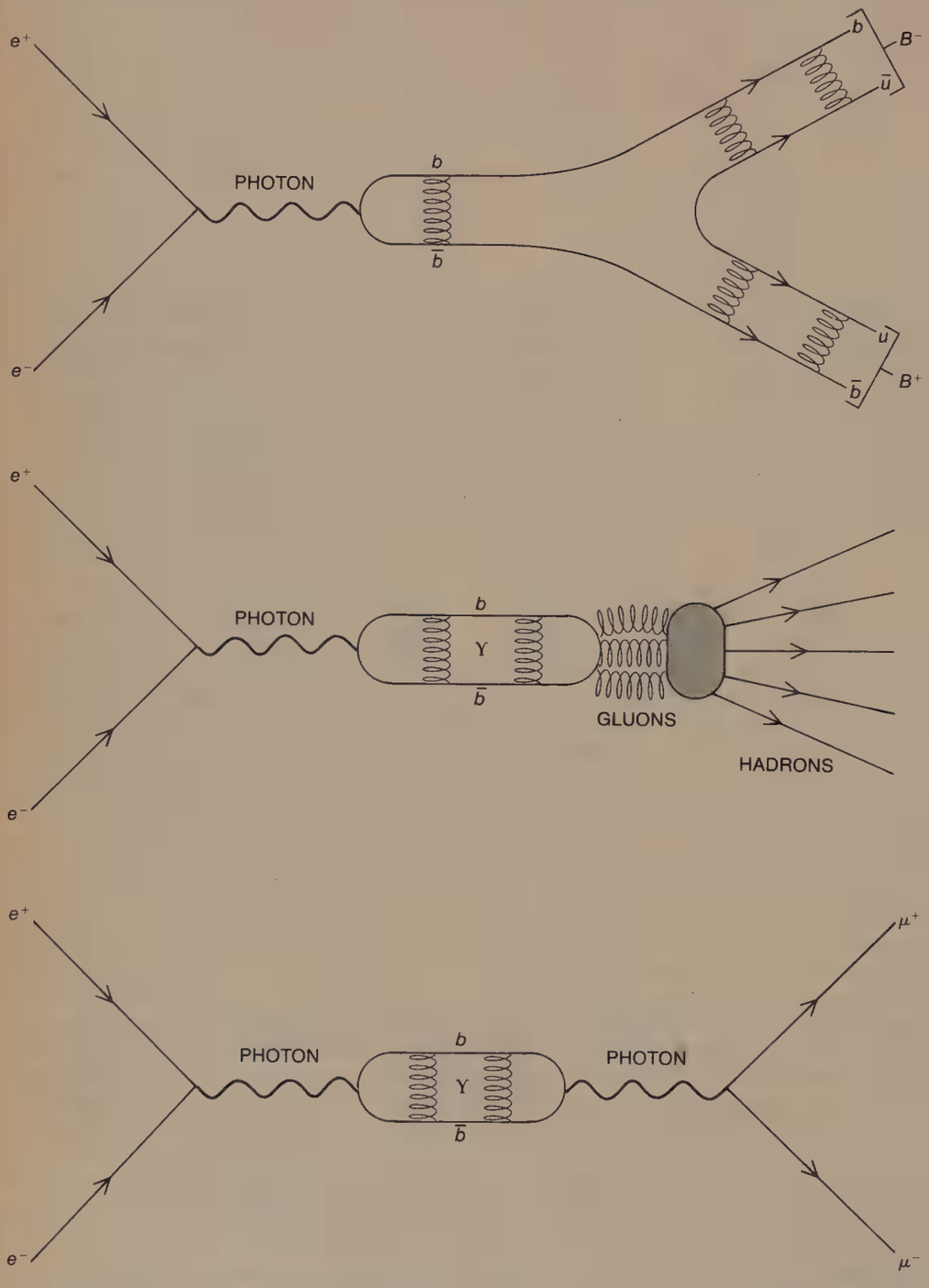
The collision technique by which the upsilon family was discovered was not well suited to the detailed study of these particles. Upsilons are rarely created in proton-nucleus collisions, and when they are, they are usually accompanied by many other particles, making it difficult to follow the process. A far better technique for studying the upsilon family, and the b quark in general, is one in which beams of electrons and positrons are made to collide with each other. Sometimes an electron and a positron will annihilate each other to form a photon, which can materialize into a quark-antiquark pair, such as $b\bar{b}$. When the energy of the electron and that of the positron add up to the mass of an upsilon particle, there is an enhanced probability of the formation of a $b\bar{b}$ pair, which can bind to form an upsilon. Since upsilons decay primarily into hadrons, the experimental technique is to vary the energy of the colliding electron-positron beams and to measure the rate of formation of hadrons. This technique was previously exploited at the electron-positron storage ring SPEAR, a facility of the Stanford Linear Accelerator Center (SLAC), in the discovery of the psi meson.

An experiment to search for the upsilons was carried out in 1978 with the electron-positron storage ring DORIS at the Deutsches Elektronen-Synchrotron (DESY) in Hamburg. Plots of the rate of hadron formation as a function of energy showed two sharp spikes at about 10 GeV, confirming the earlier indications that there were two or more distinct upsilon states. The DORIS machine had been strained to its limit in order to reach the energy of the second upsilon state, however, and it could go no higher. For this reason the third possible upsilon state was out of reach and, of greater importance, so were all $b\bar{b}$ systems in which the b and $\bar{b}$ were not tightly bound.

It was now crucial to see whether

there is indeed a third bound state of $b\bar{b}$ and, at a slightly higher energy, another "resonance" where the b and $\bar{b}$ are only loosely bound and are produced with enough energy to separate and appear as beauty-flavored particles. Theorists had already predicted the energy levels of the bound $b\bar{b}$ system based on evidence obtained from the study of the $c\bar{c}$ system of the psi mesons. It was now up to the experimentalists to confirm or refute that picture and to study the behavior of the b quark through its decays and interactions.

At the time there was a growing sense of excitement among the physicists working on the preparation of the Cornell Electron Storage Ring (CESR). This machine, built under the direction of Boyce D. McDaniel and Maury Tigner, was designed in 1975 before the existence of the upsilon was known. It was a stroke of good fortune that the upsilon system lay right in the middle of the energy range planned for the CESR. Since the CESR was first turned on in April, 1979, it has been the site of most of the detailed work done on the b quark.



BEAUTY QUARKS ARE CREATED in electron-positron collisions when the photon resulting from the annihilation of the two colliding particles materializes into a $b\bar{b}$ pair. If there is enough energy, an additional quark-antiquark pair can be created and the b and $\bar{b}$ quarks can appear in combination with the other quarks as beauty-flavored mesons (*top diagram*). At certain lower energies, where it is impossible to produce the two mesons, the $b\bar{b}$ pair is created as a bound system: the upsilon meson. Since the quarks cannot separate, they annihilate each other in one of two ways. The first type of annihilation is by the strong interaction with the emission of gluons; the gluons in turn give rise, by a poorly understood process, to hadrons, which are observed in the aftermath of the decay (*middle diagram*). This type of decay, although it is a strong-interaction process, is inhibited, with the result that electromagnetic decays of the upsilon become significant in comparison. In the latter type of decay the $b\bar{b}$ pair annihilates into a photon, which can then form, among other things, a muon-antimuon pair (*bottom diagram*).

At the CESR a "bunch" of more than 10^{11} electrons and another bunch of roughly the same number of positrons travel in opposite directions in a ring with a circumference of 768 meters. The bunches circulate within a toroidal vacuum chamber maintained at a pressure of 10^{-8} torr (10^{11} times less than atmospheric pressure). Bending magnets and focusing magnets confine the particles to a circular path within the vacuum chamber and maintain the size and shape of the bunches. A constant supply of power to the beams compensates for the energy that is radiated away. Because of the comparatively low density of the electrons and the positrons in the bunches, there are very few head-on collisions at each crossing of the bunches. The high rate of crossings (400,000 per second) ensures, however, that there is an adequate rate of electron-positron annihilations. In the argot of particle physicists these collisions are "events." Hadronic events, those in which the collision products include hadrons, are observed a few times per minute.

The detailed study of the production and decay of b quarks requires the observation and study of hundreds of thousands of hadronic events. The rate at which these events happen is low, and so it is up to the experimenters to make sure their detectors are capable of recording every interesting collision. Furthermore, the detector must not miss many particles created in the events, because such particles may be crucial to understanding what happened in the production and decay of new particles.

A great deal of effort has been invested in the design and construction of systems for detecting and studying the products of such interactions. The challenges are formidable. Events may produce many particles, both charged and neutral, that can project in all directions. Furthermore, the rarity of interesting events makes high efficiency essential. The detectors must measure and record particles that pass through the detection zone in a few billionths of a second. At the CESR the electron and positron beams cross at two diametrically opposite intersection regions. The detectors surrounding the respective crossing points reflect two different approaches to the design of detection systems for such colliding-beam experiments.

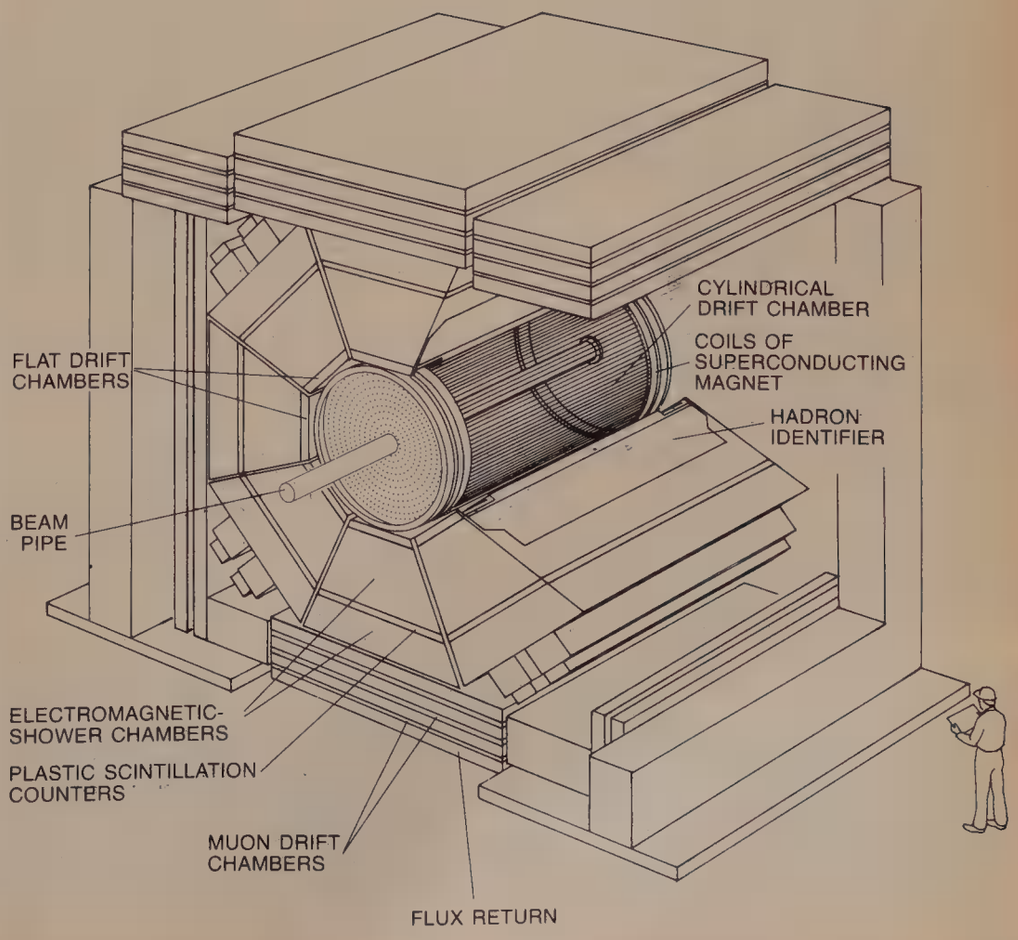
The CLEO detector, which occupies the south interaction region at the CESR, is a general-purpose detector built and operated by a group of 75 physicists (including the three of us) from seven institutions: Cornell University, Harvard University, Ohio State University, the University of Rochester, Rutgers Uni-

versity, Syracuse University and Vanderbilt University. It was designed to provide as much information as possible about each event. That includes detecting charged particles and measuring their momentum, measuring the energy and direction of photons and identifying specific kinds of particles, such as pions, kaons (*s*-flavored mesons), protons, electrons and muons. Because particles can emerge in any direction, it was necessary to surround the interaction point as completely as possible with detector components. As a result CLEO is a very large device, occupying a space six meters long, eight meters wide and nine meters high, and weighing almost 1.5 million kilograms (1,500 metric tons). Careful computer analysis of the thousands of measurements done by individual components makes it possible to develop a clear understanding of what happens in specific annihilation events.

At the heart of CLEO is a superconducting solenoid magnet one meter in radius and three meters long. Inside the magnet is a cylindrical detector, called a drift chamber, that tracks charged particles as they emerge from the collision. As many as 17 separate measurements of the position of a charged particle are recorded, enabling the experimeters to trace the particle's trajectory. By measuring the curvature of the path in the magnetic field the momentum of the particle can be determined. The process of track detection and reconstruction is a crucial element for almost all work with CLEO.

Outside the coil of the solenoid magnet are components that serve to identify different kinds of charged particles. There is a set of plastic scintillation counters, which emit a brief flash of light when a charged particle passes through them. A precise measurement





SITE OF THE DISCOVERY of beauty-flavored mesons is depicted at two scales in this pair of drawings. The plan view at the top shows the overall layout of the CESR facility. Electrons and positrons are accelerated first in a linear accelerator and then in a synchrotron ring before being injected into the storage ring. The particles are confined to separate "bunches," which circulate in opposite directions and cross at two diametrically opposed interaction points, where detectors have been installed to study the products of the electron-positron annihilations. The three-dimensional view at the bottom shows the detailed structure of the CLEO detector. A superconducting solenoid magnet sets up a strong magnetic field in the central cylindrical chamber, where the momentum of each charged particle is determined from the curvature of its trajectory. Other sensing devices are mounted in eight identical modules surrounding the central magnetic field region and in several outlying pieces of equipment.

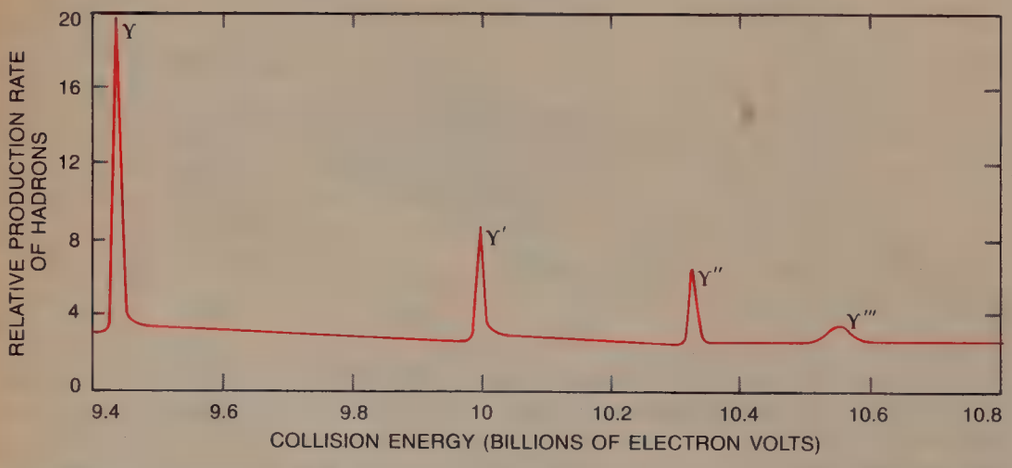
of the time of the flash, combined with accurate knowledge of the time at which the beams cross, makes it possible to calculate the time needed for a particle to travel from the collision to the counter. From this interval the speed of the particle can be determined. The measurement of the speed combined with the momentum measured with the central drift chamber leads to the determination of the particle's mass. On this basis it is possible to differentiate pions, kaons and protons.

In the original configuration of the CLEO detector the speed of a particle was also measured by a set of Cerenkov counters, which detected the light emitted when a particle passing through them exceeded the velocity of light in the gas with which the counters are filled. (The particle in this case acts somewhat like an airplane breaking the sound barrier.) The Cerenkov counters were useful for identifying electrons. They have now been replaced by gas-filled detectors that measure the rate of energy loss of each charged particle as it travels through the gas. The information can be combined with the measured momentum of the particle to yield the particle's velocity and hence its mass. In this way protons, pions, kaons and electrons can be independently identified.

Electrons have their identity confirmed in yet another part of the detector. When an electron passes through matter, it loses energy by producing what is called an electromagnetic shower. The process begins when a fast electron is deflected by a nucleus and emits an energetic photon. When the photon passes close to another nucleus, it gives rise to an electron-positron pair. The electron and the positron in turn emit photons, and the process continues with the production of a shower of secondary electrons, positrons and photons until all the energy is absorbed. The CLEO electromagnetic-shower detector is constructed of alternate layers of lead (to provide a medium for the production of showers) and charged-particle detectors (to sample the energy and extent of the shower as it develops). The shape and the energy of the showers associated with tracks coming from the collision point are analyzed to identify electrons. Showers that are not associated with any track are recognized as photons.

The detection of muons is based on their ability to penetrate large thicknesses of material. A hadron passing through matter is likely to interact strongly, and an electron will generate a shower; a muon, however, loses energy very slowly and can travel through much material without stopping. To exploit this property CLEO is surrounded by a steel absorber about a meter thick. Covering the outside of the steel are flat drift chambers that detect all charged particles emerging from the absorber. A charged particle that is observed to penetrate the steel is identified as a muon.

The CUSB detector in the north intersection region of the CESR was designed with a specific purpose in mind. It was built primarily as a high-resolution photon detector by a group of physicists from Columbia University, Louisiana State University, the Max Planck Institute for Physics and Astrophysics at Munich and the State University of New York at Stony Brook. Its main component is a set of shower detectors consisting of sodium iodide crystals. With these detectors it is possible to make precise measurements of the energy of photons. CUSB is particularly well suited to the measurement of photon energies in the range that is expected to be characteristic of transitions among the upsilon states. The CUSB photon detectors also serve to identify electron showers. Close to the collision point CUSB has tracking chambers to detect charged particles. Parts of the detector are surrounded by steel absorbers and also by counters for detecting muons.



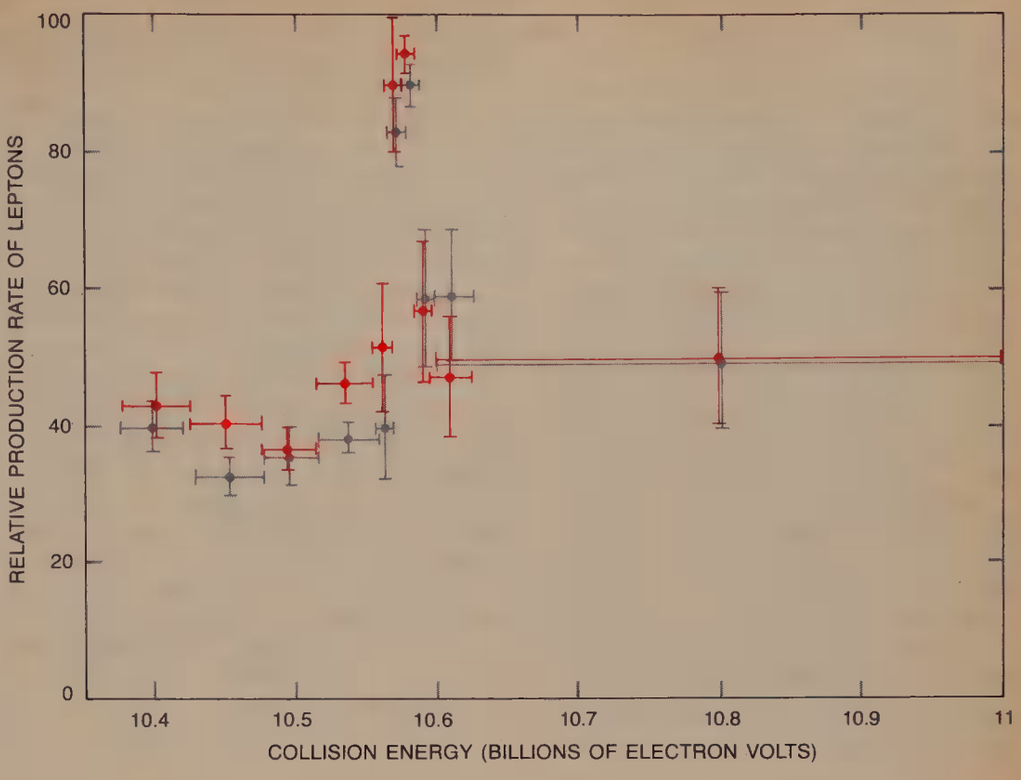
UPSILON FAMILY OF RESONANCES, representing several discrete energy states of the $b\bar{b}$ meson, are responsible for the peaks in this plot of the overall rate of hadron production as a function of the energy of the colliding electron-positron beams recorded in the CLEO detector. The Υ, Υ′ and Υ″ peaks are quite narrow, indicating that they correspond to tightly bound $b\bar{b}$ states created at energies below the threshold for the production of beauty-flavored mesons. The Υ‴ peak is wider than the other three peaks, indicating that the $b\bar{b}$ system at the higher energy state must be decaying to yield a $B\bar{B}$ pair (*see illustration on opposite page*).

The study of *b*-quark physics at the CESR began in late 1979. The upsilon-particle resonances Υ, Υ′ and Υ″ were quickly found at both CLEO and CUSB by measuring the rate of hadron production as a function of energy. The three resonances appeared as narrow peaks at the respective energies of 9.46, 10.02 and 10.35 GeV. This was the first time the Υ″ was clearly observed as a separate resonance, an event that was announced to the world of high-energy physics in a Christmas card. Since then the resonances have been carefully studied through the observation of several hundred thousand events. Some transitions among the energy states have been observed, and as rarer transitions are identified it should be possible to map out the complete spectroscopy of the upsilon system.

Since the upsilon resonances Υ, Υ′ and Υ″ are hadrons with both beauty and antibeauty, the flavors cancel and the hadrons are flavorless. Accordingly these hadrons are said to have hidden beauty. When such a hadron decays, it does so by the mutual annihilation of the *b* and $\bar{b}$ quarks, and in that case the decays of the *b* quarks themselves cannot be examined. At electron-positron energies greater than the equivalent mass of the upsilon states $b\bar{b}$ pairs can be produced with enough energy for the two quarks not to be bound, allowing them to move away from each other. As they separate, the strong force between the quarks causes a light quark-antiquark pair ($u\bar{u}$ or $d\bar{d}$) to be created. The quarks group together to form a *B* meson and a $\bar{B}$ meson ($b\bar{u}$ and $\bar{b}u$, or $b\bar{d}$ and $\bar{b}d$). In the *B* meson the beauty is no longer hidden. It is this particle that is said to display naked beauty (or bare bottom). In contrast to hidden beauty, naked beauty must decay by the weak interaction. Therefore by studying the beauty-flavored *B* mesons one can study the weak interactions of the *b* quark.

The upsilon resonances were originally interpreted as being three different energy states of the bound $b\bar{b}$ system. It seemed reasonable to expect that other resonances would exist above the threshold for decay into beauty-flavored hadrons. A bump-hunting expedition in the energy region above the Υ″ early in 1980 yielded the expected resonance, at 10.58 GeV. Unlike the first three upsilon resonances, which were all extremely narrow, the fourth resonance, the Υ‴, was quite wide, indicating that it was decaying by the unsuppressed route into *B* and $\bar{B}$; in other words, it was a source of beauty-flavored particles.

To establish clearly that this interpretation was the correct one it was necessary to find evidence for particles carrying the beauty flavor in Υ''' decays. As we have seen, quarks of one flavor can decay into those of another flavor only through the weak interaction. One class of unmistakable weak decays consists of "semileptonic" decays in which a meson decays into a muon or an electron, together with a neutrino and one or more hadrons. CLEO was well equipped for identifying both muons and electrons, and the photon detectors of CUSB were also suited to the identification of electrons; accordingly both groups of investigators set about measuring the rate of single-lepton production in Υ''' decays in order to confirm the presence of the weak decays of beauty-flavored particles.



CLEARER INDICATION that beauty-flavored particles are produced in Υ''' decays is given by these plots of the rate of lepton production in electron-positron collisions as a function of the collision energy. The sharp increases in the production rate of electrons (*gray*) and muons (*color*) appear at an energy equivalent to the mass of the Υ''' resonance. The Υ''' evidently decays into some combination of products that in turn decay into a lepton and other particles. Such "semileptonic" decays are a telltale sign of a weak-interaction process and confirm that the Υ''' resonance decays to yield a $B\bar{B}$ pair. Symbols give range of experimental error.

Because of the difficulty of detecting leptons and the low rate at which Υ''' events were accumulated, it was some time before a definitive conclusion could be drawn from the lepton yields. Through the first half of 1980, however, it gradually became clear that the rate of lepton production for events at the Υ''' energy is significantly greater than it is for events at neighboring energies. Before accepting the conclusion that we were seeing beauty-flavored hadrons, however, we had to consider the possibility that the leptons are produced by some mechanism other than the weak decay of B mesons. The observed distribution of lepton momenta is qualitatively what one would expect from the decay of a heavy meson such as the B and not from the decay of a lighter charmed meson. Since the leptons are produced singly and not in oppositely charged pairs, they cannot be the product of an electromagnetic process. When all alternative explanations were ruled out, the conclusion became clear: we were seeing naked beauty.

Although the lepton yields demonstrated the presence of naked beauty in Υ''' decays, we knew little about the B meson. Since the Υ'' is below the threshold for $B\bar{B}$ decay and the Υ''' is above it, the mass of the B meson must lie between half of the mass of the Υ'' and half of the mass of the Υ''', or in the range from 5.18 to 5.29 GeV. To demonstrate that the B meson is a real object and to measure its mass, it is necessary to reconstruct one from the debris of an Υ''' event. To do so calls for taking the tracks of an event in various combinations and determining whether any particular combination can be demonstrated to be the decay products of a B meson. The procedure is difficult because a B meson often decays into a large number of particles, some of which are neutral and can escape detection. Others emerge along the beam pipe and are not seen for that reason. What is more, each event has both a B and a $\bar{B}$ meson, and so the number of incorrect combinations can be extremely large.

By concentrating on decay modes of the B that involve only a few particles the CLEO group has recently succeeded in reconstructing B mesons. To do so a sample of 150,000 events collected over a nine-month period was searched for the charmed particles known as neutral D mesons that decayed to form a kaon-pion pair. In some cases the D^0 was produced directly in the decay of a B meson. In others the B decayed into a charged excited state of the charmed meson, which in turn decayed into a D^0 and a charged pion. The reconstructed B's had all decayed into a D^0 or a charged excited state and one or two charged pions, giving a total of from three to five charged particles and no neutral ones. This is in marked contrast to the typical B, which decays into nine particles, three of which are neutral. The masses of the 18 B candidates cluster around 5.274 GeV.

Our mass measurement for the B demonstrates that the Υ''', at 10.58 GeV, is only .03 GeV above the threshold for $B\bar{B}$ production (twice the B mass, or 10.55 GeV). When the Υ''' resonance decays into $B\bar{B}$, there is just enough energy to make up the mass of the two mesons; as a result they are produced virtually at rest and with no accompanying particles to confuse the experimenter. Thus the Υ''' is an ideal "factory" for making beauty-flavored particles: it is both an abundant source of B mesons and an uncluttered place to observe and study the decays of b quarks. There are many features of these decays we should like to explore. For instance, it is important to learn how quickly the b quark decays and what it decays into. If history is any guide, these experiments will lead to an improved understanding of the relations among the different flavors of quarks.

Although the b quark is expected to decay slowly on the time scale of elementary-particle processes, the decay seems very fast on a macroscopic time scale. The lifetime of the b is predicted to be between 10^{-13} and 10^{-14} second. In that brief interval a B meson traveling at a speed of 10 percent of the speed of light will move only .0002 centimeter. This distance is too small to be measured with the CLEO detector. So far the best that can be said from the CLEO data is that the lifetime is less than 10^{-10} second. A similar study done at the electron-positron storage ring named PETRA at DESY had the advantage of higher-energy (and therefore faster-moving) B's. In spite of an extremely low rate of B production, that experiment gave an up-

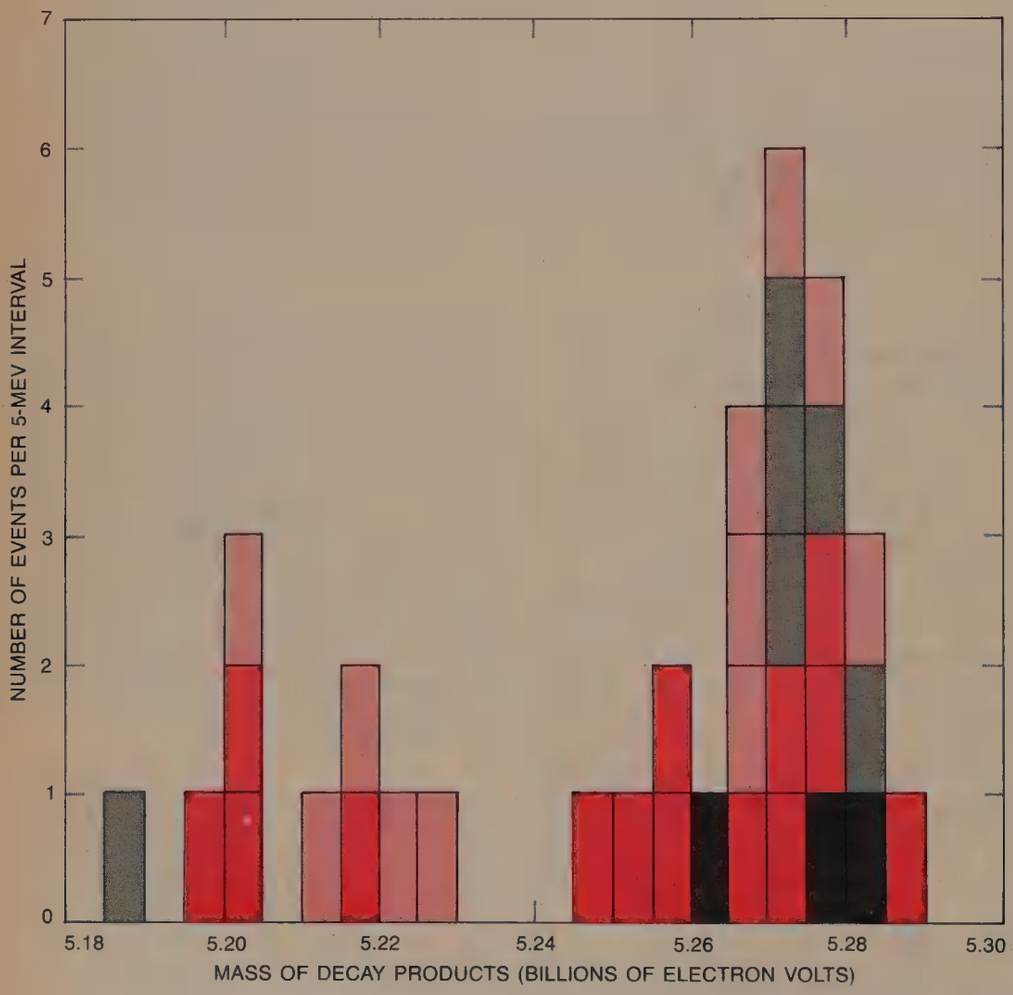
per limit on the B lifetime of 5×10^{-12} second.

What are the products of the decay of the b quark? The possibilities can be grouped into three classes. The standard view is that the b emits a W^- and is transformed into a c or a u. An alternative is for the b to emit a Z^0, transforming the b into an s or a d. This is a beauty-changing process involving a Z^0 and is forbidden if the b is part of a doublet with t. Finally, it is possible that the b decays in some entirely new and unexpected way, not involving the emission of either a W^- or a Z^0. Let us consider each of these possibilities in turn.

In the standard view, with the b decaying into a W^- and a u or a c, the end products of b decay must include the particles into which the W^- itself decays. The options are $e^-\bar{\nu}_e$, $\mu^-\bar{\nu}_\mu$, $\tau^-\bar{\nu}_\tau$, $\bar{u}d$ and $\bar{c}s$. In general the quark options are three times likelier than the lepton options. By counting the possible outcomes and allowing for some necessary theoretical adjustments one can predict that the end products of b decay will include an electron about 13 percent of the time and a muon the same percentage of the time. The prediction agrees quite well with experimental measurements from CLEO and CUSB, lending support to the standard view of b decay.

This theoretical picture makes no prediction about whether the b quark decays more often to c or to u. Accordingly that is one of the most important things one can hope to learn about the b quark. The recent direct observations of D^0 mesons in B-meson decay demonstrates that the b quark does in fact decay to c at least some of the time. There are two procedures by which the decay preference of the b quark has been measured. The well-known dominance of the decay of c into s suggests the most obvious approach: one would expect B decays to produce more strange particles (kaons) if the b quark decays predominantly into c than if b decays into u. CLEO can detect both charged and neutral kaons. A careful study of kaon production shows there is a definite enhancement in the rate of kaon production in Υ''' events compared with nonresonance events, which translates to a combined rate for charged and neutral kaon production of about 1.4 particles per B decay. Theoretically one would expect about .8 kaon for each B decay if b always decays into u and 1.6 if b always decays into c. The measurement suggests that the decay of b into c is dominant.

A better way to examine the question is to analyze the momentum distribution of the leptons from semileptonic B decays. Among the products of one of these decays there must be a lepton, a neutrino and one or more hadrons. If b decays into c, a charmed particle must be created. The lightest charmed particle is the D meson, with a mass of 1.87 GeV. Since this mass is quite large, it limits the amount of energy that can be carried away by the lepton as momentum. In contrast, if b decays into u, it is not necessary to create such a massive hadron, and it is possible for the lepton to have more momentum. By examining the production of high-momentum leptons and comparing these observations with what would be expected if b decays into c (or if b decays into u) it has been possible to demonstrate again that b decays primarily into c. In fact, it appears from this evidence that b decays into u no more than 5 percent of the time.



RECONSTRUCTION OF BEAUTY-FLAVORED MESONS was accomplished by the CLEO group by examining the charged-particle tracks from the decay of the Υ''' resonance and tabulating them in various combinations according to four different models of the subsequent decay of a beauty-flavored meson (*key at left*). The resulting bar chart shows a peak in the vicinity of 5.274 billion electron volts. The specific decay modes considered are unusual in that the B meson decays into a small number of charged particles and no uncharged particles (other than the extremely short-lived D^0, or neutral charmed meson). The rarity of these decays and the difficulty of filtering them from other Υ''' decay processes is reflected in the fact that the small number of entries shown is the result of nine months of data collection, totaling 150,000 collision events. The abbreviation MeV in the vertical scale stands for millions of electron volts.

The standard view of the decay of the b quark, which has been the framework for the foregoing discussion, demands that the b quark share a doublet with the t quark. The discovery of the t quark is crucial to the confirmation of this model. The t quark has been sought in experiments at the electron-positron storage rings PEP (at SLAC) and PETRA (at DESY). The failure to find the t may simply mean that the maximum energy of these searches (37 GeV) is too low to allow the production of bound $t\bar{t}$ states. The alternative is that the standard theory is incorrect and another theoretical description must be found.

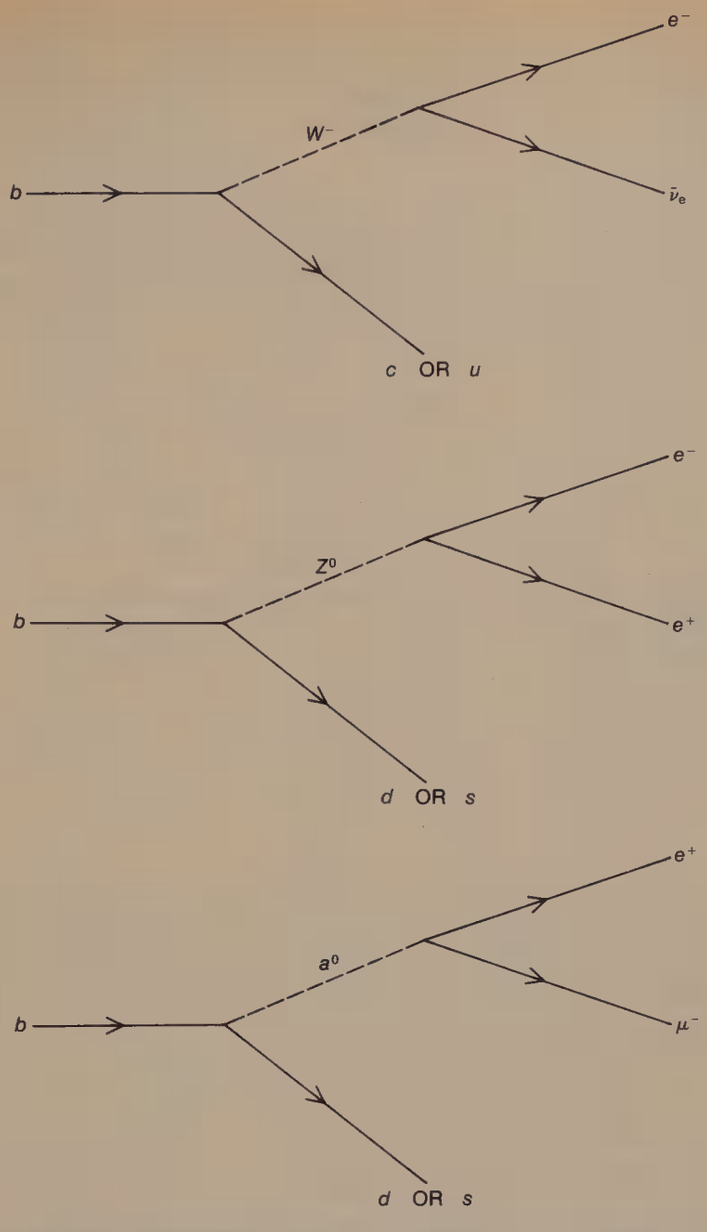
If the standard theory is wrong and the b quark does not have a partner, one must consider the possibility that it decays by emitting a Z^0. (Remember, this process is not allowed in the standard theory.) Independent of whether the b turns into a d or an s, the decay products of the Z^0 must be present among the end products of b decay. The Z^0 can produce e^+e^-, $\mu^+\mu^-$, $\tau^+\tau^-$, $\nu\bar{\nu}$, $u\bar{u}$, $d\bar{d}$, $s\bar{s}$ or $c\bar{c}$. Here the e^+e^- and $\mu^+\mu^-$ options are very attractive because they yield clear signals. A search for indications of two routes of b decay, $b \rightarrow \mu^+ + \mu^- +$ hadrons and $b \rightarrow e^+ + e^- +$ hadrons, has yielded no evidence of such

decays, leading to the conclusion that the Z^0 is not a major contributor to b decay.

Finally, there is the possibility that the b decays neither by W^- emission nor by Z^0 emission but in some entirely new way. Within the limits of rather general theoretical considerations only two additional kinds of decay are possible. The first is for the b to decay into a quark and two leptons from different doublets. The second is for the b to decay into a lepton and two antiquarks. If the b quark decays by the first of these exotic processes, with two leptons per decay, one would expect events with many muons, electrons, taus and neutrinos in various combinations. The taus in turn would decay, releasing part of their energy to neutrinos, which would be undetectable. If this kind of decay were significant, one would expect to observe frequent electron and muon production and considerable missing energy in B-decay events. By simultaneously measuring the yield of muons and electrons and the detectable energy in each Υ''' event CLEO has shown that this kind of decay is not dominant.

In the second class of exotic b decays three more antiquarks than quarks are created (one that was originally bound to the b quark and two that were created in its decay). The presence of the three antiquarks results in the creation of one antibaryon per B-meson decay. Similarly, the excess of quarks in $\bar{b}$ decay results in one baryon per $\bar{B}$ decay. Although there are many kinds of baryons (and antibaryons) that can be produced, all eventually decay into protons or neutrons (or into antiprotons or antineutrons). In CLEO protons and antiprotons can be identified, but neutrons and antineutrons escape undetected. If the b decays into a lepton and two antiquarks, one would expect B-decay events to yield, say, a muon, an electron, an antiproton or a large amount of missing energy. By simultaneously measuring the yields of muons, electrons and antiprotons, along with the detected energy per event, the CLEO experimenters have shown that this second kind of exotic decay is not dominant either.

In sum, all the evidence supports the view that the b quark decays by emitting a W^-. Furthermore, there are indications that most of the time it turns into a c quark rather than a u quark.



SEVERAL POSSIBILITIES for the decay of the b quark were investigated by the CLEO group. In the standard theoretical model, based on the assumption of three quark doublets, the b quark decays into either a c quark or a u quark by emitting a W^- boson (*top diagram*). The W^- can then decay into five possible final states, two of which ($e^-\bar{\nu}_e$ and $\mu^-\bar{\nu}_\mu$) are the semileptonic modes that were observed in the discovery of naked beauty. If the t quark, the b quark's hypothetical partner, does not exist, there must be a second mechanism for b decay, namely one in which the b quark decays into a d quark or an s quark by emitting a Z^0 (*middle diagram*). Again there is a range of possible final states, two of which (e^+e^- and $\mu^+\mu^-$) have been shown by the CLEO group to occur rarely, if at all. Exotic decay processes, not involving the usual gauge bosons, have been suggested by some theorists engaged in exploring alternatives to the standard model. One example is the decay of the b quark into a d quark or an s quark by the emission of a^0, an exotic neutral boson (*bottom diagram*). The a^0 (for which there is no experimental evidence) need not obey the rules that apply to conventional gauge bosons and so can decay into two leptons of different types, such as a negative muon and a positron. Experiments with CLEO have shown that such exotic decay processes also are uncommon.

Although an understanding of the b quark and its interactions is developing, a number of open questions remain. What is the lifetime of the b quark? At present only upper limits have been established; they are far above the expected lifetime, however, and therefore provide no new information for evaluating the theory. What fraction of the time does b decay into u rather than into c? Data from CLEO and CUSB show that the fraction of b decays into u is small, but how small is it? Does the t quark exist? By showing that beauty-changing processes involving a Z^0 occur infrequently (if at all) and by showing that the unexpected processes that involve neither W^- nor Z^0 are also rare (if not absent) the CLEO experiments have ruled out almost all theories that can avoid having a t quark. Perhaps the t quark has not been found yet because it is too massive, and searches with higher-energy colliding-beam machines will find it. If the t does not exist, the theory of quark decays is in deep trouble.

Even if it is assumed that the t quark will ultimately be found, there remain some profound questions about the quark and lepton doublets. A perfectly satisfactory world can evidently be built of one lepton doublet (e^-, ν_e) and one quark doublet (u, d). Yet at least three doublets of each type are thought to exist, and more may still be discovered. Why does nature behave in such an uneconomical way? How are the various doublets related and how do they differ? What does "flavor" really mean? Further study of the b quark may help to answer some of these questions.

Creole Languages

These widely scattered languages show striking similarities. The development of Creole in Hawaii suggests children learn a language by first constructing an abstract form of a creole

by Derek Bickerton

The ancient Greek historian Herodotus records the story of Psamtik I, pharaoh of Egypt in the seventh century B.C., who set out to discover the original language of humanity. On royal decree two infants were taken away from their parents and put in the care of a mute shepherd, who was instructed to raise the children in isolation from other people. The shepherd was to take note of the first word uttered by the children; "uncorrupted" by the language of their forefathers, Psamtik reasoned, they would begin to speak in the pure tongue from which all other languages were derived. The first intelligible sound the children made was "bekos," which meant bread in the ancient language Phrygian. Therefore, Psamtik maintained, the original language of humanity is Phrygian.

The story has amused generations of linguistics students. Most linguists, who have taken it for granted that no such experiment should ever be carried out, have dismissed the Psamtik experiment as being defective in design and unlikely to yield any useful result. Indeed, the assumption that an "original" vocabulary can be recovered is overoptimistic, and linguistic isolation of the individual, which has been documented in a few cases of severe child abuse, usually results in the absence of language. Nevertheless, a modified form of the experiment has been repeated many times over the past 500 years among the children of slaves and laborers who were pressed into service by the European colonial powers.

These laborers, who were shipped from many parts of the world to tend and harvest crops in Africa, the Indian Ocean region, the Orient, the Caribbean and Hawaii, were obliged to communicate within their polyglot community by means of the rudimentary speech system called pidgin. Pidgin speech is extremely impoverished in syntax and vocabulary, but for the children born into the colonial community it was the only common language available. From these modest beginnings new native languages evolved among the children, which are generically called creole languages. It can be shown that they exhibit the complexity, nuance and expressive power universally found in the more established languages of the world.

Taken at face value, the development of many different creole languages suggests that the search for a single, original language is misguided. For many years, however, scholars have noted a remarkable similarity of structure among all the creole languages. It can now be demonstrated, by considering the origin of creole language in Hawaii, that similarities among creoles cannot be accounted for by contact with other languages, either indigenous or imported. The finding suggests that what is common to creole languages may indeed form the basis of the acquisition of language by children everywhere. There is now an impressive body of evidence to support this hypothesis: between the ages of two and four the child born into a community of linguistically competent adults speaks a variety of language whose structure bears a deep resemblance to the structure of creole languages. Hence, by an ironic stroke of justice, the surviving linguistic remnants of colonialism may offer indispensable keys to the study of our own linguistic heritage.

The historical conditions that favored the development of creole languages are well known. Between 1500 and 1900 England, France, the Netherlands, Portugal and Spain established numerous labor-intensive, agricultural economies on isolated littorals and underpopulated tropical islands throughout the world. The colonies were engaged primarily in monoculture, usually sugar, and their economic viability depended on an abundance of cheap labor imported from distant regions under conditions of chattel slavery. Workers were drawn first from West Africa and later from East Africa, India and the Orient, and they spoke a variety of mutually incomprehensible languages.

Under more salutary conditions of immigration the workers or their children would eventually have learned the language of the local colonial power, but two factors combined to keep them from doing so. First, the number of speakers of the colonial languages rarely exceeded 20 percent of the total population, and it was often less than 10 percent. In other words, there were relatively few people from whom the dominant language could have been learned. Second, the colonial societies were small, autocratic and rigidly stratified. There were few chances for prolonged linguistic contact between field laborers and speakers of the dominant language.

Except in Hawaii, there is little reliable documentary evidence concerning the early linguistic history of the colonial societies. It has generally been assumed that pidgin developed as a contact language solely to allow communication between masters and workers and among workers from various immigrant groups. Creole languages then arose among the children of the workers through the "expansion" of pidgin; there was little occasion for the children to use the ancestral languages of their parents, and they still lacked access to the language of the dominant culture. What is meant by the term "expansion" has remained obscure until my colleagues and I began our studies in Hawaii.

The unique advantage for the study of creole language in Hawaii is that the details of its formation can be reconstructed at least in part from the speech of people still living. Although Hawaiian contact with Europeans goes back to 1778, it was not until 1876 that a revision in the U.S. tariff laws, allowing the free importation of Hawaiian sugar, enabled Hawaiian sugar plantations to increase their output by several hundred percent. A polyglot force of indentured laborers, made up of Chinese, Filipinos, Japanese, Koreans, Portuguese, Puerto Ricans and others, began to be assembled, and by 1900 it outnumbered the other groups in Hawaii, both native and European, by a ratio of two to one.

A pidgin based on the Polynesian lan-

guage Hawaiian initially served as a means of communication between immigrants and the locally born, but the annexation of Hawaii by the U.S. in 1898 eventually led to the replacement of Hawaiian by English. After 1900 the Hawaiian language declined, and pidgin Hawaiian was replaced as a lingua franca by a pidgin based on English. By the time we began our intensive study of language variation in Hawaii in the early 1970's there were still many survivors, both immigrant and locally born, from the years 1900 until 1920.

Our recordings of locally born people make it clear that the process of creolization was under way by 1900 and was certainly complete by 1920. Most of the linguistic features that characterize Hawaiian Creole English are present in the speech of working-class people born in Hawaii since 1905; before that date the proportion of Creole speakers to the rest of the population falls off rapidly. On the other hand, the speech of immigrants is always some form of pidgin, although just what form it takes depends on the date of the immigrant's arrival in Hawaii as well as the immigrant's language background. The pidgin spoken by the earliest immigrants among our subjects is much more rudimentary than that spoken by the later ones, probably because the latter were exposed to Creole as well as pidgin. Nevertheless, the distinction between pidgin and Creole remains fundamental: anyone familiar with Hawaii can quickly identify the ethnic origins of any immigrant on the basis of speech patterns alone. Without a conversational topic or a person's physical appearance as a guide, however, no one can reliably identify the ethnic origins of any locally born speaker solely on the basis of the speaker's pronunciation or the grammatical structure of the utterances.

One of the main characteristics of pidgin, therefore, is its variability from speaker to speaker. Each immigrant seems to have gone about the task of inventing a makeshift language in some individual way. For example, pidgin speakers of Japanese ancestry generally place the verb at the end of a sentence, as in "The poor people all potato eat" ("All that the poor people ate were potatoes"). Filipino pidgin, however, places the verb before the subject: "Work hard these people" ("These people work hard"). More often word order follows no fixed principle except the pragmatic one that old, shared information is stated near the beginning of a sentence and new information near the end.

It is probably the case that anything expressible in Creole, or in English for that matter, can also be expressed in pidgin. Nevertheless, the pidgin speaker is at a great disadvantage, because pidgin lacks many of the building blocks possessed by all native languages. Such everyday necessities of language as articles, prepositions and auxiliary verbs are either absent or appear sporadically in a quite unpredictable fashion. Pidgin sentences have no subordinate clauses, and single-clause utterances frequently lack verbs.

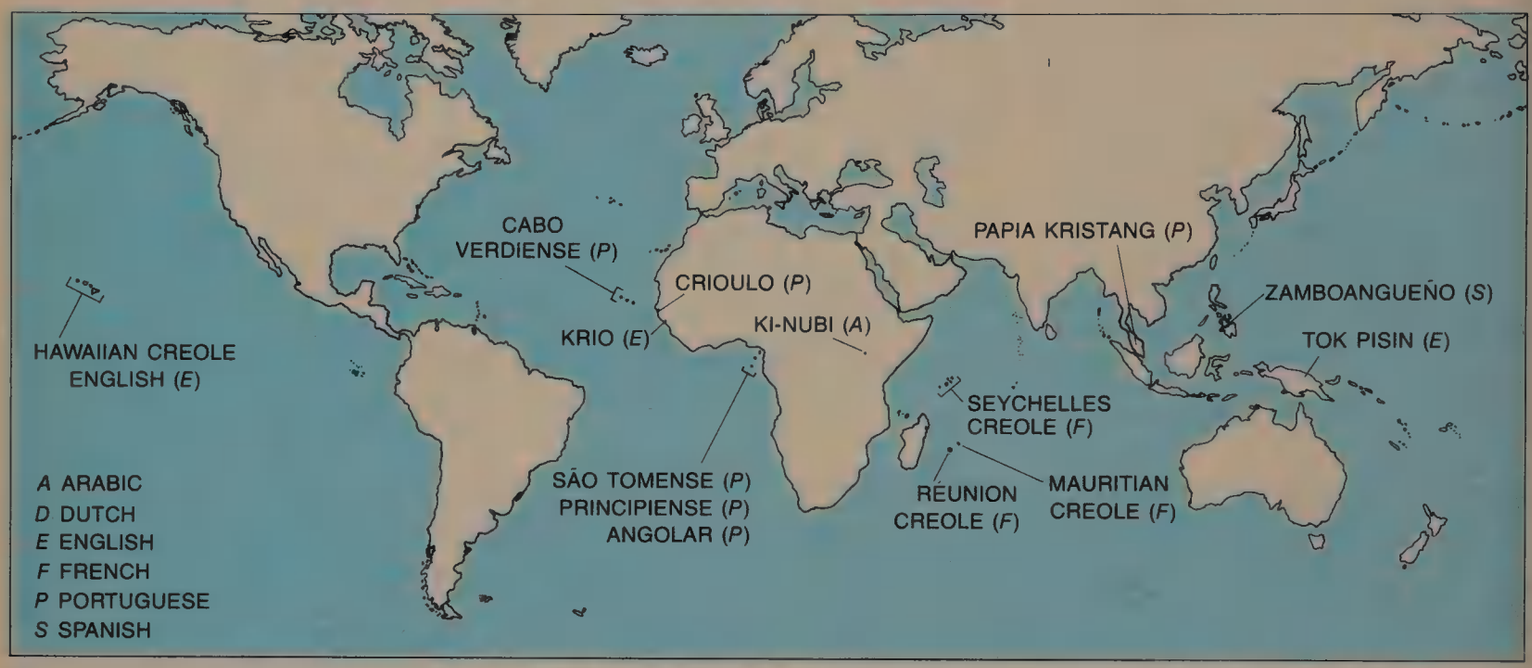
The first of the following examples

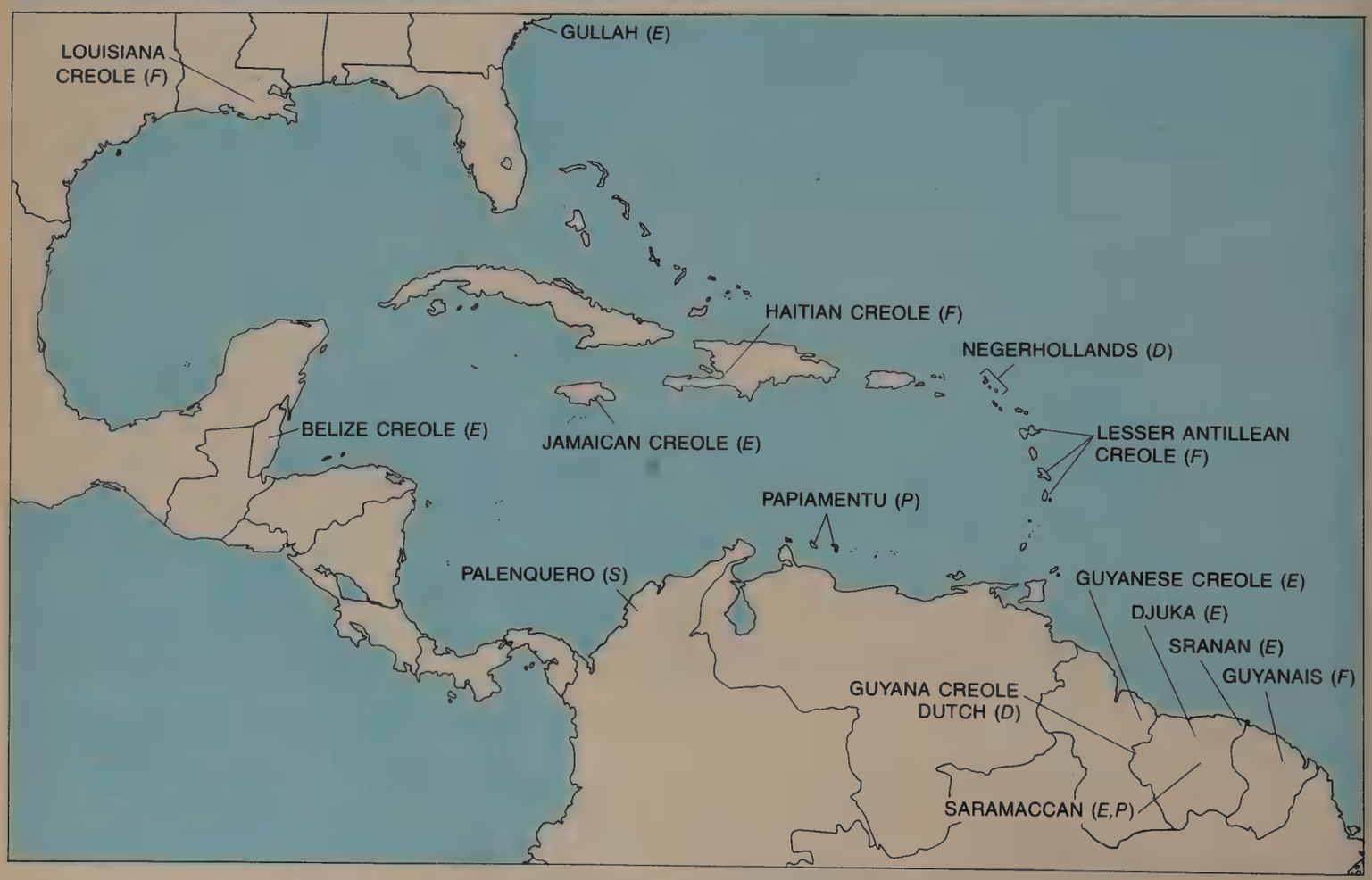


INDENTURED SUGARCANE WORKERS, who spoke a rudimentary language called pidgin, are shown in a photograph made in Hawaii by the late Ray Jerome Baker in 1924. Thousands of such workers from many countries were brought to Hawaii in the late 19th and early 20th centuries to meet the labor demands of large sugarcane and pineapple plantations. Pidgin language developed out of the need for communication among the various language groups within this polyglot labor force; the workers in the photograph, for example, although primarily from the Philippines, originally spoke a variety of mutually incomprehensible languages such as Visayan, Ilocano and Tagalog. The children of pidgin-speaking parents remained largely isolated from speakers of Hawaiian or English, and they inherited no consistent linguistic model from the previous generation that could meet their needs. There arose among these children a language much more sophisticated than pidgin called Hawaiian Creole English, whose vocabulary is primarily English but whose structure, or grammar, is quite distinct from that of English or any other noncreole language. Socioeconomic circumstances similar to those in Hawaii frequently gave rise to pidgin languages throughout the world; these languages were later developed into creole languages by the children of immigrant laborers. The original photograph is in the collection of Ray Jerome Baker and Robert E. Van Dyke in Honolulu.

was recorded from a pidgin-speaking Korean; omitted words are bracketed in the translation: "And a too much children, small children, house money pay" ("And [I had] too many children, small children, [I had] to pay the rent"). The second example was recorded from a Japanese speaker: "Before mill no more Filipino no nothing" ("Before the mill [was built, there were] no Filipinos here at all"). The third example, recorded from the speech of a retired bus driver, illustrates the heroic measures needed to say anything out of the ordinary in pidgin: "Sometime good road get, sometime, all same bend get, enguru [angle] get, no? Any kind same. All same human life, all same" ("Sometimes there's a good road, sometimes there's, like, bends, corners, right? Everything's like that. Human life's just like that").

The language-learning task confronted by the child born into a community of such speakers is far different from the task imposed on the child who is surrounded by linguistically competent adults. The children of English or Chinese parents, for example, are presented with accurate models to follow. Although their mistakes are seldom overtly corrected, they can almost constantly check their own utterances against those of older speakers and adapt them





WORLDWIDE DISTRIBUTION of creole languages reflects the historical circumstances of their development. Almost all creoles arose on isolated tropical littorals or islands, where colonial powers had established agricultural economies based on cheap immigrant labor. The geographic dispersion of the colonies suggests that creole languages developed independently of one another. The letters in parentheses after the name of each language indicate the colonial language from which most of the vocabulary of the creole is borrowed.

where necessary. When they have mastered the simpler structures of their language, more complex structures are readily available.

For the Hawaiian-born child of immigrant parents, however, there was no consistent linguistic model for the basic word order of simple sentences and often no model at all for the more complicated structures of language. Many such children were born of interethnic or interracial marriages, and so even at home there was little occasion to speak the native language of either parent. Moreover, even among the children not born of linguistically mixed parents there was considerable incentive to abandon the parents' native language and adopt some version of pidgin in the company of peers and neighboring adults. Like first-generation immigrant children elsewhere, the children of Hawaiian immigrants often became bilingual or even trilingual, and they adopted the common language of their peers as a native language in spite of considerable efforts by their parents to maintain the ancestral tongue.

The historical evidence is consistent with the view that the structure of Creole arose without significant borrowing from other languages. Bilingual or trilingual children of school age need not (and usually do not) mix up the structural features of the languages they speak, and there is no reason to suppose such crossovers were common in Hawaii. The most compelling argument for the autonomous emergence of Creole, however, is its observed uniformity. How, within a single generation, did such a consistent and uniform language develop out of the linguistic free-for-all that was pidgin in Hawaii? Even if all the children of various immigrant groups had begun by learning the languages of their parents, and even if the differences among the various pidgins had been smoothed by interaction and contact among the children, the homogeneity of the language that developed remains in need of explanation. Fifty years of contact among pidgin-speaking adults were not enough to erase the differences among the national language groups; the homogeneity must have resulted from the differences between children and adults.

One might still suppose the structural uniformity of Creole is derived from certain structures of one of the ancestral languages or perhaps from certain structures of English, the language of the plantation owners. There are numerous differences, however, between the structure of Creole and the structure of any of the languages with which Creole speakers might have been in contact. In English, for example, it is possible to refer to an object or a group of objects in a nonspecific way, but English grammar forces the speaker to state in advance whether the number of unspecified objects is one or many, singular or plural. One must say either "I am going to the store to buy a shirt" or "I am going to the store to buy shirts," even though one may not want to commit oneself in advance to buying any particular number of shirts.

In Creole a grammatically neutral marker for number can be employed on the noun "shirt" in order to avoid specifying number: "I stay go da store for buy shirt" ("I am going to the store to buy shirt"). Moreover, in Creole the addition of a definite or an indefinite article to "shirt" changes the meaning of the sentence. In saying "I stay go da store for buy one shirt" the Creole speaker asserts the shirt is a specific one; in the sentence "I stay go da store for buy da shirt" the speaker further presupposes

PIDGIN	HAWAIIAN CREOLE ENGLISH
Building—high place—wall part—time—now-time—and then—now temperature every time give you.	Get one [There is an] electric sign high up on da wall of da building show you what time an' temperature get [it is] right now.
Now days, ah, house, ah, inside, washi clothes machine get, no? Before time, ah, no more, see? And then pipe no more, water pipe no more.	Those days bin get [there were] no more washing machine, no more pipe water like get [there is] inside house nowadays, ah?
No, the men, ah—pau [finished] work—they go, make garden. Plant this, ah, cabbage, like that. Plant potato, like that. And then—all that one—all right, sit down. Make lilly bit story.	When work pau [is finished] da guys they stay go make [are going to make] garden for plant potato an' cabbage an' after little while they go sit down talk story ["shoot the breeze"].
Good, this one. Kaukau [food] any kind this one. Pilipin island no good. No more money...	Hawaii more better than Philippines, over here get [there is] plenty kaukau [food], over there no can, bra [brother], you no more money for buy kaukau [food], 'a'swhy [that's why].

PIDGIN AND CREOLE versions of identical sentences illustrate the structural differences between pidgin and Creole in Hawaii. Pidgin, which is spoken only by immigrants, varies widely from speaker to speaker. Although one can probably say anything in pidgin that can be said in English or Creole, the structure of pidgin is extremely rudimentary. Pidgin sentences are little more than strings of nouns, verbs and adjectives, often arranged to place old, shared information first and new information later in the sentence. Creole arose in Hawaii only among the children of immigrants, and it is much richer in grammatical structure than pidgin. Moreover, the rules of Creole grammar are uniform from speaker to speaker, and they resemble the structural rules of other creoles. English versions of words and phrases are given in brackets.

ENGLISH	HAWAIIAN CREOLE ENGLISH
The two of us had a hard time raising dogs.	Us two bin get hard time raising dog.
John and his friends are stealing the food.	John-them stay cockroach the kaukau.
He doesn't want to play because he's lazy.	He lazy, 'a'swhy he no like play.
How do you expect to finish your house?	How you expect for make pau you house?
It would have been better if I'd gone to Honolulu to buy it.	More better I bin go Honolulu for buy om.
The one who falls first is the loser.	Who go down first is loser.
The man who was going to lay the vinyl had quoted me a price.	The guy gon' lay the vinyl bin quote me price.
There was a woman who had three daughters.	Bin get one wahine she get three daughter.
She can't go because she hasn't any money.	She no can go, she no more money, 'a'swhy.

STRUCTURAL DIFFERENCES between sentences in Hawaiian Creole and their English equivalents show that the grammar of Creole did not originate as a grammar borrowed from English. For example, the past perfect tense of a verb in Creole is expressed by the particles "bin" or "wen," which precede the main verb, instead of by the suffix "-ed." Nonpunctual, or progressive, aspect is expressed by the word "stay" instead of by the suffix "-ing." In the English sentence "The two of us had a hard time raising dogs" the rules of grammar oblige the speaker to indicate that the noun "dog" is either singular or plural. In the Creole version of the sentence, however, neither singular nor plural is implied. There are also relatively insignificant lexical differences between the two languages: "cockroach" is picturesquely employed as a verb, and "kaukau," which may be derived from the Chinese pidgin term "chowchow," is a common word for "food." Equally striking structural differences are found between Hawaiian Creole and other languages, such as Chinese, Hawaiian, Japanese, Korean, Portuguese, Spanish or the Philippine languages, with which speakers of Hawaiian Creole might have been in contact.

that the listener is already familiar with the shirt the speaker is going to buy.

There are many other features of Creole that distinguish it from English. Whereas in English there is a past tense, which is usually marked with the suffix "-ed," in Creole there is a tense called the anterior tense, which is marked with "bin" for older speakers and with "wen" for younger speakers. The anterior tense is somewhat like the English past perfect: "had walked" in English is "bin walk" in Creole, and "walked" in English is simply "walk" in Creole. In order to distinguish irreal, or possible, actions or processes from actual ones, English employs the conditional or the future tense. In Creole all such irreal circumstances are expressed by the particle "go," which is placed before the main verb and marks what linguists call modality. For example, the sentence "If I had a car, I would drive home" is rendered in Creole as "If I bin get car, I go drive home."

There is also a Creole auxiliary verb that marks what linguists call aspect; it too is placed before the main verb and indicates that the action expressed by the verb is nonpunctual, or in other words repeated, habitual, continuing or incomplete. In order to say "I run in Kapiolani Park every evening" in Creole one must say "I stay run in Kapiolani Park every evening." If the particle "stay" is omitted by the Creole speaker, the action is understood to be completed or nonrepetitive.

In English there is no straightforward way to distinguish purposes that have been accomplished from those that have not. The sentence "John went to Honolulu to see Mary" does not specify whether or not John actually saw Mary. In Creole grammar the ambiguity must be resolved. If John saw Mary and the Creole speaker knows that John saw Mary, the speaker must say, "John bin go Honolulu go see Mary." If John did not see Mary or if the speaker does not know whether or not John saw Mary, the speaker must say, "John bin go Honolulu for see Mary."

Similar distinctions could be drawn between the grammatical structure of Creole and the structure of other contact languages, such as Hawaiian, Ilocano (the language spoken in the north of the Philippine island of Luzon) and Japanese. There are also resemblances, but most of them are confined to idiomatic expressions. For example, the Creole expression "O the pretty," which means "How pretty he [she/it] is," is a literal translation of the Hawaiian-language idiom "O ka nani." In the main, however, our investigations strongly suggest that the basic structures of Creole differ from those of other languages. Although it might seem that some children of immigrants could transfer the structures of their parents' native languages onto the evolving Creole language, they did not do so. The structural linguistic input that was available to the children was apparently not used in the development of Creole.

Even if it could be demonstrated that all the grammatical structures of Creole were borrowed, cafeteria-style, from one contact language or another, the uniformity of Creole would present a difficult question: How did the speakers who invented Creole come to agree on which structure to borrow from which language? Without such agreement Creole could not be as uniform as it is. Yet it seems highly implausible that the agreement could have been reached so quickly. If there had been massive borrowing from ancestral languages, differences in the version of Creole spoken by various groups would have persisted at least one generation beyond the first generation of speakers.

There is another dimension to the problem of the uniformity of Hawaiian Creole. It turns out that creole languages throughout the world exhibit the same uniformity and even the same grammatical structures that are observed in Hawaii. The finding is all the more remarkable when it is compared with the rather poor correspondence in structure I have noted between Hawaiian Creole and other contact languages in Hawaii. For example, the distinction made in Hawaiian Creole between singular, plural and neutral number is also made in all other creole languages. Similarly, in all other creole languages there are three invariant particles that act as auxiliary verbs and play the roles that "bin," "go" and "stay" play in Hawaiian Creole.

In Haitian Creole, for example, the word "té" marks the anterior tense of the verb, the word "av(a)" marks irreal modality and the word "ap" marks the aspect of the verb as nonpunctual. Thus in Haitian Creole the phrase "I have been walking" is rendered "m [I] t'ap [té + ap] maché." Similarly, in Sranan, an English-based creole found in Surinam (formerly Netherlands Guiana), the anterior tense marker is "ben," the irreal modality marker is "sa" and the nonpunctual aspect marker is "e." The phrase "He would have been walking" is rendered "A [he] ben sa e waka." Most important, there is a strict order that must be followed in all creole languages when more than one of these markers is present in a sentence. The particle for tense precedes the particle for modality, and the particle for modality precedes the particle for aspect.

Finally, consider the grammatical distinction I have noted between purposes accomplished and unaccomplished. The same distinction, absent in English, is found in all creoles. In Mauritian Creole, a creole based on the French vocabulary that is used on the island of Mauritius, a sentence such as "He decided to eat meat" can be expressed in two ways. If the subject of the sentence carried out his decision, the sentence is rendered "Li ti desid al mâz lavian," which literally means "He decided go eat meat." If the decision was not carried out, the sentence is rendered as "Li ti desid pu mâz lavian," or literally "He decided for eat meat." In Jamaican Creole the sentence "He went to wash" must be rendered either as "Im gaan fi bied" ("He went with the intention of washing") or as "Im gaan go bied" ("He went to wash and completed the task").

These examples only suggest the extent of the structural similarities among creole languages. The similarities seem unaffected by the wide geographic dispersion of the creoles and the variation among the languages such as Dutch, English and French from which they draw the greatest part of their vocabulary. Scholars such as Hugo Schuchardt began to point out the resemblance in the 19th century, and in the 1960's many examples were explored in detail by Douglas Taylor, by Robert Wallace Thompson of the University of the West Indies and by Keith Whinnom of the University of Exeter. Thus even before the development of Hawaiian Creole was reasonably well understood the grammatical similarities among the creole languages of the world were recognized as an important finding that required explanation.

The linguist's first reaction to such a finding is to look for a common ancestor of the similar languages. For ex-

VERB FORM
BASE FORM ("HE WALKED"; "HE LOVES")
ANTERIOR ("HE HAD WALKED"; "HE LOVED")
IRREAL ("HE WILL/WOULD WALK"; "HE WILL/WOULD
NONPUNCTUAL ("HE IS/WAS WALKING")
ANTERIOR + IRREAL ("HE WOULD HAVE WAL "HE WOULD HAVE LOVED")
ANTERIOR + NONPUNCTUAL ("HE WAS/HAD BEEN WALKING")
IRREAL + NONPUNCTUAL ("HE WILL/WOULD BE WALKING")
ANTERIOR + IRREAL + NONPUNCTUAL ("HE WOULD HAVE BEEN WALKING")

CONJUGATION OF THE VERB is similar in all creole languages, in spite of superficial lexical differences. Moreover, the creole system is quite distinct from the one encountered in English and in most other languages. The table gives conjugations in Hawaiian Creole, Haitian Creole and Sranan (an English-based creole spoken in Surinam, the former Nether-

ample, it has been conjectured that the linguistic ancestor was a contact language that grew out of Portuguese and certain West African languages in the course of the first Portuguese explorations of Africa in the 15th and 16th centuries. According to the hypothesis, this contact language was subsequently spread around the world by Portuguese sailors, changing its vocabulary but not its syntax or semantics as it entered the sphere of influence of another colonial power. Superficially such an explanation might seem to be consistent with the development of Creole in Hawaii, because Portuguese laborers were brought to the islands in large numbers during the late 19th and early 20th centuries.

There are several serious flaws in the account. First, Hawaiian Creole bears scant resemblance to any of the contact languages, including Portuguese. Second, the claims of linguistic similarity between creoles and Portuguese or between creoles and West African languages are grossly exaggerated. Most important, our study of hundreds of Hawaiian speakers has made it clear that Hawaiian Creole almost certainly originated in Hawaii. We found no surviving immigrant who speaks anything approximating a creole language; instead every immigrant we surveyed speaks some variety of pidgin. If Hawaiian Creole was primarily an imported language, it would have been carried by immigrants, and presumably it would have been learned by others among the immigrant population. One must therefore conclude that Hawaiian Creole arose among the children of immigrants, where it is now found. Moreover, if a creole language could develop in Hawaii without ancestry, it can arise anywhere else in a similar way.

The implications of these findings are far-reaching. Because the grammatical structures of creole languages are more similar to one another than they are to the structures of any other language, it is reasonable to suppose most if not all creoles were invented by the children of pidgin-speaking immigrants. Moreover, since creoles must have been invented in isolation, it is likely that some general ability, common to all people, is responsible for the linguistic similarities.

The suggestion that people are biologically predisposed to use language is not a new one: for more than two decades Noam Chomsky of the Massachusetts Institute of Technology has argued that there is an innate universal grammar underlying all human languages. The universal grammar is postulated largely on the grounds that only by its means could children acquire a system as enormously complex as a human language in the short time they do. Studies by the late Eric H. Lenneberg tend to confirm Chomsky's hypothesis. The acquisition of language resembles the acquisition of other complex and flexible aspects of the child's behavior, such as walking, which are undoubtedly controlled to some degree by neurophysiological development. The universal grammar conjectured by Chomsky is a computing device, somehow realized neurologically, that makes a wide range of grammatical models available to the child. According to Chomsky, the child must then "select" which of the available grammatical models matches the grammar of the language into which the child is born.

The evidence from creole languages suggests that first-language acquisition is mediated by an innate device of a rather different kind. Instead of making a range of grammatical models available, the device provides the child with a single and fairly specific grammatical model. It was only in pidgin-speaking communities, where there was no grammatical model that could compete with the child's innate grammar, that the innate grammatical model was not eventually suppressed. The innate grammar was then clothed in whatever vocabulary was locally available and gave rise to the creole languages heard today.

The implications of this hypothesis call into question an idea that most linguists, including Chomsky, have tacitly accepted for many years, namely that no one of the world's languages is easier or harder for the child to acquire than any other. If there is a creole grammar somehow imprinted in the mind, creole languages should be easier to acquire than other languages. How is it, then, that not all children grow up speaking a creole language? The answer is they do their best to do just that. People around them, however, persist in speaking English or French or some other language, and so the child must modify the grammar of the native creole until it conforms to that of the local language.

Two kinds of linguistic evidence are relevant for testing the hypothesis. First, if some grammatical structure of creole is at variance with the corresponding grammatical structure of the local language, one should find that children

NONSTATIVE VERBS			STATIVE VERBS		
WAIIAN CREOLE	HAITIAN CREOLE	SRANAN	HAWAIIAN CREOLE	HAITIAN CREOLE	SRANAN
WALK	LI MACHÉ	A WAKA	HE LOVE	LI RÊMÊ	A LOBI
IN WALK	LI TÉ MACHÉ	A BEN WAKA	HE BIN LOVE	LI TÉ RÊMÊ	A BEN LOBI
GO WALK	L'AV(A) MACHÉ	A SA WAKA	HE GO LOVE	L'AV(A) RÊMÊ	A SA LOBI
TAY WALK	L'AP MACHÉ	A E WAKA	—	—	—
BIN GO WALK	LI T'AV(A) MACHÉ	A BEN SA WAKA	HE BIN GO LOVE	LI T'AV(A) RÊMÊ	A BEN SA LOBI
BIN STAY WALK	LI T'AP MACHÉ	A BEN E WAKA	—	—	—
GO STAY WALK	L'AV AP MACHÉ	A SA E WAKA	—	—	—
BIN GO STAY WALK	LI T'AV AP MACHÉ	A BEN SA E WAKA	—	—	—

lands Guiana) for stative and nonstative verbs. Stative verbs are verbs such as "like," "want" and "love," which cannot form the nonpunctual aspect; in English, for example, one cannot add "-ing" to a finite stative verb. The base form of the verb refers to the present for stative verbs and to the past for nonstative verbs. The anterior tense is roughly equivalent to the English past tense for stative verbs and to the English past perfect tense for nonstative verbs. The irreal mode includes the English future, conditional and subjunctive. In all the creole languages the anterior particle precedes the irreal particle, and the irreal particle precedes the nonpunctual particle. In Hawaiian Creole, however, "He bin go walk" has come to mean "He walked" instead of "He would have walked," and the forms "He bin stay walk," "He go stay walk" and "He bin go stay walk," although they were widespread before World War II, are now almost extinct because of the growing influence of English in Hawaii. The bracketed English translations are provided only as a rough guide to the meaning.

CHILD LANGUAGE	ENGLISH CREOLES
Where I can put it?	Where I can put om? (Hawaii)
Daddy throw the nother rock.	Daddy t'row one neda rock'tone. (Jamaica)
I go full Angela bucket.	I go full Angela bucket. (Guyana)
Lookit a boy play ball.	Luku one boy a play ball. (Jamaica)
Nobody don't like me.	Nobody no like me. (Guyana)
I no like do that.	I no like do that. (Hawaii)
Johnny big more than me.	Johnny big more than me. (Jamaica)
Let Daddy get pen write it.	Make Daddy get pen write am. (Guyana)
I more better than Johnny.	I more better than Johnny. (Hawaii)

SENTENCES SPOKEN BY CHILDREN between two and four years old, all born of English-speaking parents, are strikingly similar to sentences in English-based creole languages. The similarities among creole languages and the likelihood that the languages arose independently of one another suggest that creoles develop among children whenever there is no adequate native language to serve as a model. The author conjectures that if children were removed from their native English-language community at the age of about two, they would grow up speaking a language whose vocabulary would be primarily English but whose grammar would be a creole.

make systematic errors with respect to the structure of the local language. On the other hand, if the two grammatical structures tend to agree, one should find extremely early, rapid and errorless acquisition of the local-language structure.

Consider the systematic error observed by David McNeill of the University of Michigan in the speech of a four-year-old boy. In one of McNeill's observing sessions the boy complained, "Nobody don't like me," and the boy's mother responded by correcting the sentence: "Nobody likes me." The boy then repeated his sentence and the mother repeated her correction no fewer than eight times. Finally, the child altered his sentence and shouted in exasperation, "Nobody don't likes me."

The error is found in many English-speaking children between three and a half and four years old, including children who are not exposed to dialects of English that employ double negatives. There are many languages, such as French and Spanish, that also employ double negatives, but the only languages that allow negative subjects with negative verbs are creoles. For example, in Papia Kristang, the Portuguese-based creole language of the Malay Peninsula, one can say, "Angkosa nte mersimentu," or literally, "Nothing not-have value." In Guyanese Creole, which is based on English and found in Guyana (formerly British Guiana), one can say, "Non dag na bait non kyat," or literally, "No dog did not bite no cat."

A second instance of systematic error is found in the formation of children's questions. Children learning English often indicate questions only by their intonation; the subject and the auxiliary verb are almost never reversed. For example, children repeatedly say things such as "You can fix this?" even though they have heard countless questions such as "Can you fix this?" Similarly, no creole language distinguishes questions and statements on the basis of word order; the difference is marked by intonation alone.

Consider the sentence "A gon' full Angela bucket." Although such a sentence is unacceptable in English, it is perfectly acceptable in Hawaiian Creole, Guyanese Creole or any of several other creoles related to English. It is synonymous with the sentence "I'm going to fill Angela's bucket," but it differs from the structure of the English sentence in the following ways. First, the first-person pronoun "I" is reduced to "A"; second, the auxiliary verb "am" is omitted; third, the forms "go" or "gon" are used to mark the future tense; fourth, the word "to" in the infinitive is omitted; fifth, the adjective "full" is employed as if it were a transitive verb, and sixth, the possessive marker "-'s" is omitted. All these features are characteristics of creoles, but this sentence was not uttered by a creole speaker. It was spoken by the three-year-old daughter of an English-speaking linguist.

When a feature of the local language matches the structure of creole, children avoid making errors that would otherwise seem quite natural. For example, children learning English acquire the suffix "-ing," which expresses duration, at a very early age. Even before the age of two many children say things such as "I sitting high chair," where the verb expresses a continuing action. One would expect that as soon as the suffix was acquired it would be applied to every possible verb, just as the suffix "-s" that marks the English plural is frequently overgeneralized to nouns such as "foot" and "sheep."

One would therefore expect children to utter ungrammatical sentences such as "I liking Mommy" and "I wanting candy." Remarkably, such errors are almost never heard. Children seem to know implicitly that English verbs such as "like" and "want," which are called stative verbs, cannot be marked by the suffix "-ing" to indicate duration. The distinction between stative and nonstative verbs is fundamental to creole languages, however, and no marker of continuing action can be employed with a stative verb in creoles either.

The distinction between specific and nonspecific reference, which I have already discussed, is an important feature of creole languages. In English the distinction can be subtle, but young children nonetheless acquire it with ease. Michael P. Maratsos of the University of Minnesota constructed a series of sentences for children to complete, for which the completions depended on the distinction between specific and nonspecific reference. For example, the sentence "John has never read a book," which makes nonspecific reference to the noun "book," can be completed by the phrase "and he never will read a book"; it cannot be completed by the phrase "and he never will read the book." Similarly, the sentence "John read a book yesterday," in which a specific book is presupposed, can be completed by the phrase "and he enjoyed the book"; it cannot be completed by the phrase "and he enjoyed a book." Children as young as three years were able to make such distinctions correctly about 90 percent of the time.

Many more studies of language acquisition will have to be carried out before the structure of creole languages can be firmly accepted as the basis of first-language acquisition. Daniel Isaac Slobin of the University of California at Berkeley has suggested that there is a set of processes children apply to any language they hear, which he calls basic child grammar. Slobin's most recent work, which is not yet published, cites evidence from several languages for the hypothesis, and it now appears that basic child grammar and creole languages may have much in common.

If creole languages represent the manifestation of a neurologically determined program of child development, then Psamtik was by no means the fool he has been taken for. It may be possible to discover, at least in general outline, the structure of human language in the early stages of its development. Moreover, in attempting to reconstruct such a language linguists may be able to answer questions the pharaoh did not even ask: How did human language originate? What are the minimum prerequisites for such a thing as language to arise in a species? If such questions can be answered or even formulated in a precise and coherent way, we shall be much closer to understanding what makes the human species different from others.

The IBM Personal Computer. More bang for your buck.



Now you can have the best prize of all: IBM quality at a new price* that's better than ever. The system includes 64KB of user memory, one 160KB diskette drive plus a color/graphics adapter.

However you plan to use your IBM Personal Computer, you'll win big—because it can help suit your needs in the office, home or school.

Growth options? Certainly. Add a color monitor if you'd like. Or a printer. Or more memory. And you can choose from IBM's growing library of personal computer software to make your system a truly useful, valuable and versatile tool.

To get all the news about this important IBM price/performance story, ring the bell at your authorized retail dealer.

IBM®

*This price applies to IBM Product Centers. Prices may vary at other stores.
For more information on where to buy the IBM Personal Computer, call 800-447-4700. In Alaska or Hawaii, 800-447-0890.

THE AMATEUR SCIENTIST

The physics of the follow, the draw and the massé (in billiards and pool)

by Jearl Walker

Billiards and pool have the feel of physics. Balls collide with each other and the rails of the table like gas molecules in some two-dimensional tank. Actually the physics of billiards and pool is subtler. For example, a skilled player can impart spin to the ball, achieving such effects as the follow, the draw and the massé. Indeed, the interaction of the cue and the ball may be the most challenging application of classical mechanics. To master the forces and trajectories of both billiards and pool one must play the game often and analytically. A helpful step toward that objective is an understanding of the physics of the games.

Recently Todd King of Temple City, Calif., sent me his analysis of some of the classic shots in pool. Until a few years ago almost the only study concerning the dynamics of billiards was in lecture notes by Arnold Sommerfeld, who is better known for his work on early quantum mechanics. Last year David F. Griffing of Miami University devoted a chapter to pool and billiards as part of his book *The Dynamics of Sports: Why That's the Way the Ball Bounces.* These three sources inform the following discussion of the physics of billiards and pool. After describing some simple relations I shall take up a few of the famous trick shots outlined in *Byrne's Treasury of Trick Shots in Pool and Billiards,* by Robert Byrne.

When the cue strikes the cue ball, both horizontal and rotational motions are imparted. For simplicity assume that the cue stick is horizontal and delivers only a horizontal force. Although the force can be applied anywhere on the ball's surface facing the player, the shock sets the ball in horizontal motion just as if the force were applied at the center of mass.

Now assume that the stroke by the cue is in the vertical plane passing through the ball's center of mass, namely on an imaginary vertical line running through the center of the face toward the player. The location of the blow along this line has no direct bearing on either the ball's initial velocity or its momentum (the product of mass and velocity). They are set by two other factors in the collision. One factor, over which the player has virtually no control, is the duration of the collision. The second factor, easily controlled by the player, is the force on the ball. A "hard" shot generates more velocity and momentum than a "soft" shot because the force in the collision is greater.

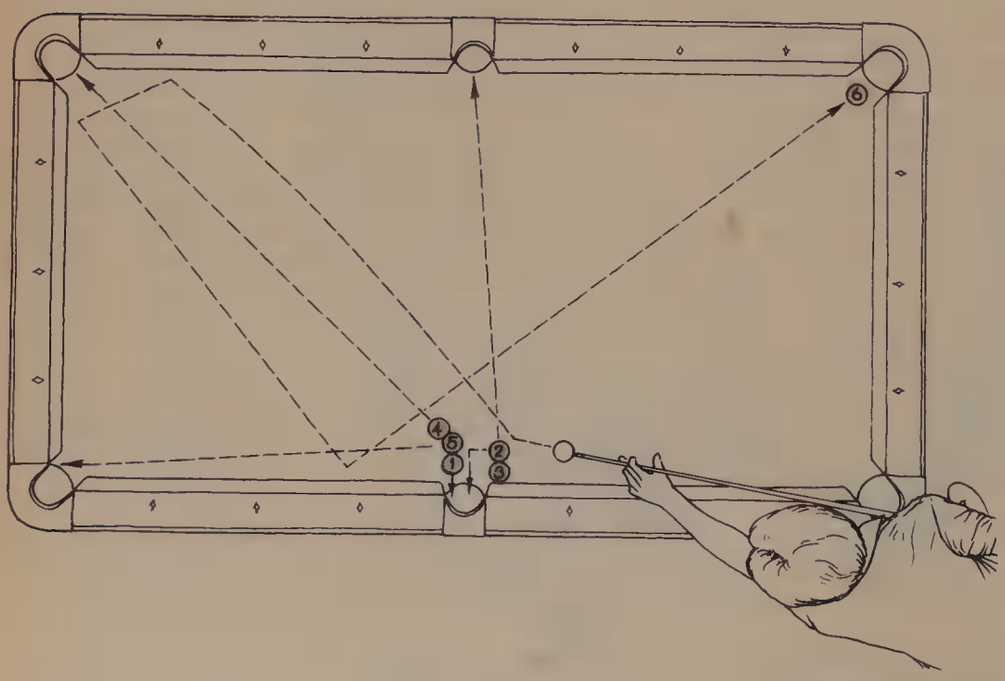
In addition to horizontal motion the cue also generates a torque that makes a ball rotate about its center of mass. The magnitude of the torque is equal to the product of the force and a lever arm that represents the vertical distance between the middle of the ball and the point where the cue strikes. The torque increases as the distance of the blow from the middle of the ball increases.

The torque determines the initial rate at which the ball spins about the center of mass. The spin is proportional to the torque divided by the ball's moment of inertia (a number that takes into account not only the mass of an object but also the distribution of the mass around the axis of rotation). For a cue ball rotating about an axis through its center of mass the moment of inertia is two-fifths of the mass multiplied by the square of the radius. The factor of 2/5, which arises from the shape of the ball, plays a role in a player's decision about where to stroke the ball in certain shots.

If the player wants an initially nonspinning ball, he should strike it at the height of its center of mass. The lever arm for such a stroke is zero, and so the torque and spin are zero. With a higher blow the collision has a lever arm and hence a measurable torque. The ball moves forward because of the force of the collision, and it spins about its center of mass because of the torque. The ball has topspin: the top of the ball moves away from the player faster than it otherwise would. Striking the ball below the center results in backspin.

The player's stroke therefore controls three features of the motion. The force determines the velocity of the ball over the table. The lever arm of the force determines the direction of spin. The product of the force and the lever arm determines the rate of spin.

Without friction from the surface of the table the cue ball would continue moving until it hit a rail or another ball. Even a surface worn smooth from play can provide significant friction, however, if the ball slips on the cloth. The friction can be high enough to alter both the horizontal and the rotational motion of the ball and thereby significantly change the shot. If the ball rolls over the table without slipping, the friction is low and affects little more than the maximum distance of roll.



The "just showin' off" shot

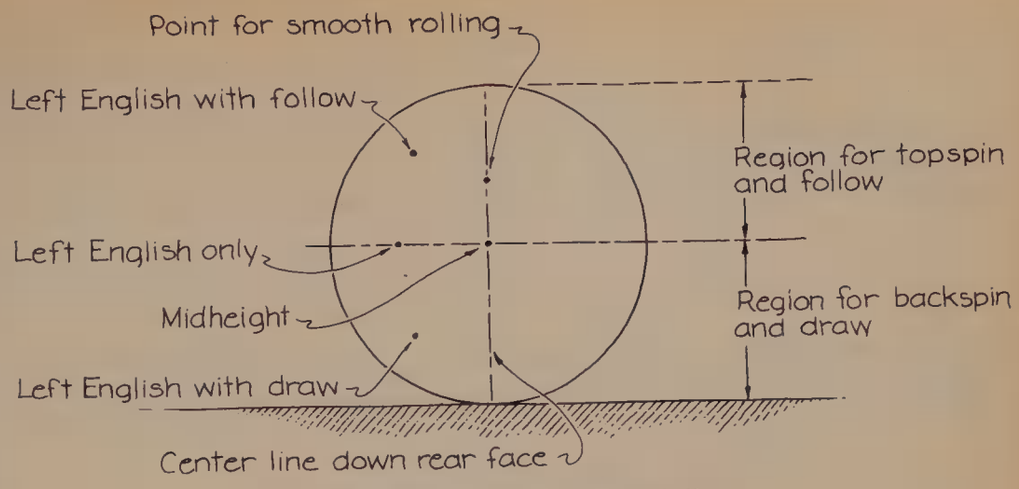
The friction on a slipping ball depends on the weight of the ball and on the surface texture of the cloth and the ball, but it is independent of the rate at which the ball slips. The direction of the force depends entirely on the direction of slip. Suppose the player delivers a large topspin to the cue ball; the bottom surface then slides toward him and the center of mass moves away. At the point of contact the friction force is opposite to the sliding. (The force is away from the player.) Since the friction opposes the sliding, it begins to decrease the spin of the ball about its center of mass. And since the friction force is away from the player, it continues to propel the ball forward and away from the player. A ball given a large topspin can run for a long time because of this additional propulsion.

Suppose the player imparts backspin. The force sends the ball away and the torque sets it into a spin that makes the bottom slide over the cloth in the same direction. The friction force on the bottom on the ball is thus toward the player. Again the friction tends to slow the spin, but this time its effect on the center of mass is rearward. As the ball slides, its forward motion and spin are slowed. Eventually the backspin is eliminated and the ball begins to roll without sliding. A ball hit with backspin runs only a short distance because the friction opposes the motion of the center of mass.

A ball with topspin will slip unless the speed of its center of mass is equal to the spin multiplied by the radius. Then the ball's forward motion is exactly matched by the motion of the bottom surface through the point of contact with the table. To achieve this match instantly the player must stroke the ball at a point that is exactly above the middle by a distance equal to two-fifths of the radius. This relation is established by the fact that the 2/5 in the formula for the moment of inertia must be canceled by the 2/5 in the formula for the lever arm of the torque.

If the cue ball is struck higher than at this special point, the spin rate is at first too large for a match to be achieved, but friction tends to force the ball into the matched state. The friction (directed forward) reduces the spin and increases the speed of the center of mass. If the ball does not run into anything, the match of speeds is eventually made and the ball begins to roll smoothly.

If the ball is struck between the middle and the special point, its spin is in the right direction for smooth rolling but is too small. At the point of contact with the cloth the ball's surface has a small rearward motion owing to the spin and a larger forward motion owing to the speed of the center of mass. The net slipping motion there is forward, creating a rearward friction force that tends to increase the spin and decrease the motion of the center of mass until the match that causes a smooth roll is achieved.



Strategic places to strike the cue ball

If the ball is struck below the middle, the spin is in the wrong direction for smooth rolling. This time the friction induced by slipping reduces both the spin and the speed of the center of mass. Eventually the spin stops and the ball begins to roll smoothly.

A skilled player can impart a long or short run to a ball by striking it at a point relative to the special point for smooth rolling. If he wants the ball to reach the far side of the table quickly, he must strike it above the special point so that the friction propels the ball.

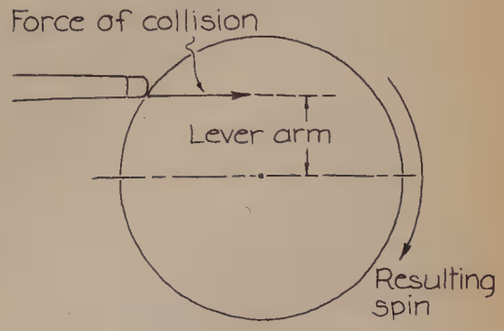
The player is more likely to be concerned with how the ball is rotating when it strikes another ball. (The other balls are called the object balls.) A collision between a cue ball and an object ball transfers momentum from the cue ball. In a head-on collision the transfer is complete, leaving the cue ball with a motionless center of mass. In a glancing collision the cue ball loses only part of its momentum and continues to travel. In any collision virtually none of the cue ball's rotation is transferred because the friction between the surfaces of the colliding balls is minute and the collision is brief. Only with significant friction could the cue ball transfer spin to an object ball.

The absence of transferred spin leads to two interesting shots. Suppose the cue ball is hit with topspin and collides head on with an object ball while it is still sliding. Just after the collision the cue ball's center of mass is motionless but the ball continues to spin. The forward-directed friction generated by the spin slows the spin and begins to propel the center of mass. Soon the cue ball begins to roll again, following the object ball. This is a follow shot. A cue ball with topspin is often said to have "follow" or "follow English."

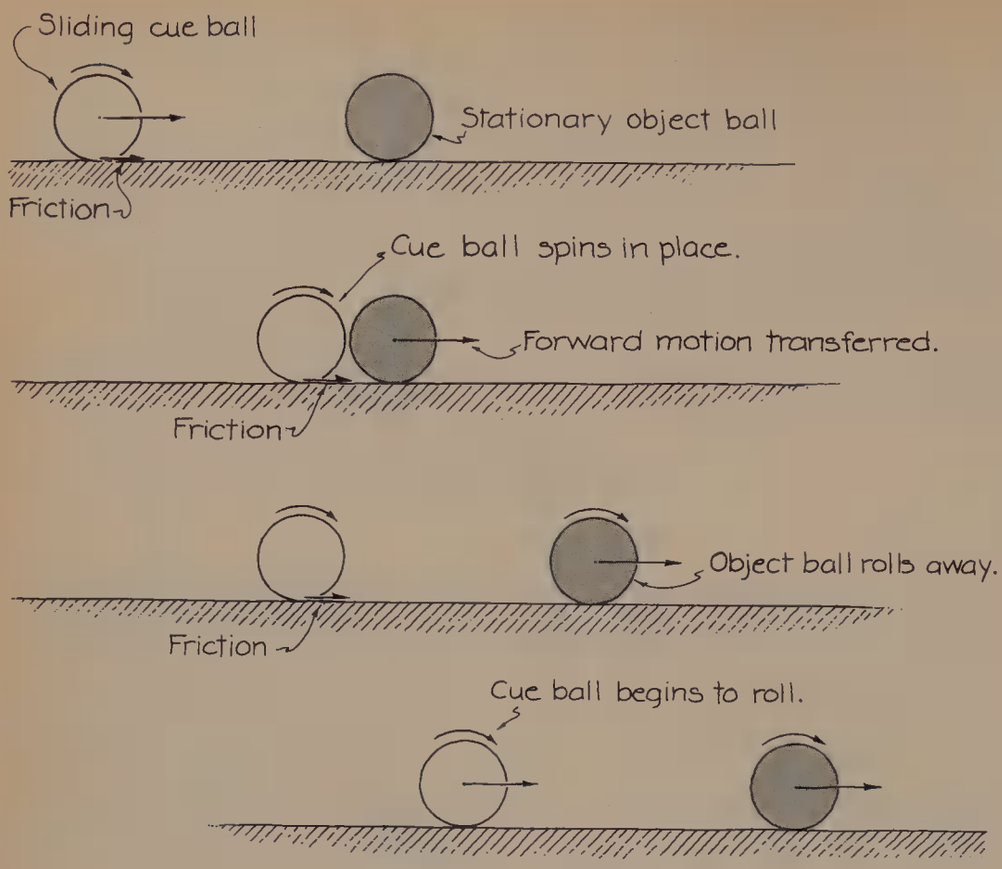
If the cue ball is given backspin, it will return to the player after hitting an object ball head on. The collision leaves the cue ball with a motionless center of mass but with the same amount of spin. The friction generated by the sliding surface is toward the player. As the friction slows the spin and propels the center of mass, the ball begins to roll smoothly toward the player. This is a draw shot. A ball with backspin is often said to have "draw" or "draw English."

A follow shot is depicted in the middle illustration on the next page. A player who wants to pocket the four ball and the seven ball with a single shot strikes the cue ball with follow, thus causing the four ball to ricochet off the seven ball and into the pocket. The collision of the cue ball with the four ball leaves the cue ball momentarily spinning in place, but sliding friction soon propels it toward the pocket again. In the meantime the seven ball has bounced off the rail near the pocket and come into line between the cue ball and the pocket. The cue ball then pockets the seven ball and comes to a stop.

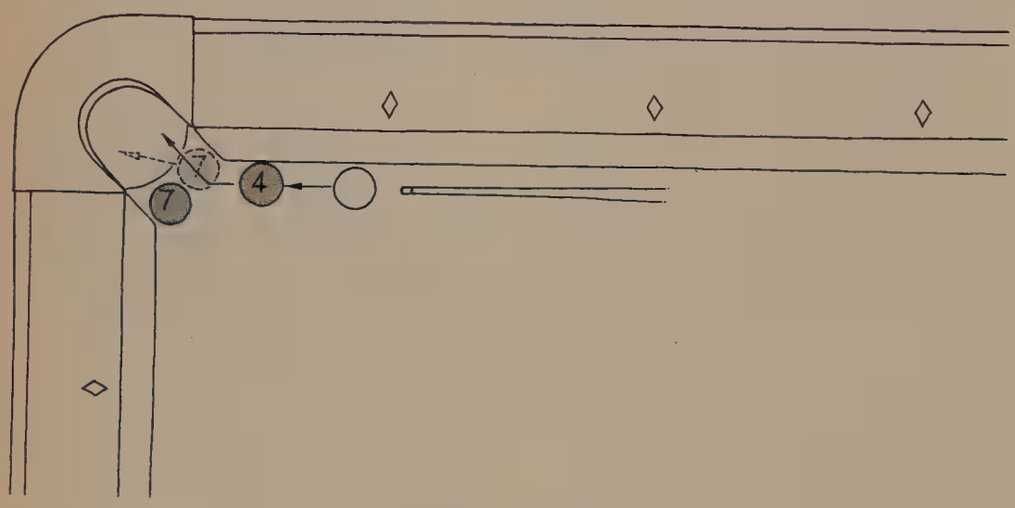
So far I have written only of stroking the cue ball along a vertical line through the center. The result is topspin or backspin about a horizontal axis. If the ball is stroked elsewhere, the axis of spin still passes through the center of mass but is no longer horizontal. A stroke on the left side of the ball is said to give left English, on the right side right English. From overhead left English is a clockwise spin about the vertical. As before, the rate of spin depends on the lever arm associated with the force. The farther



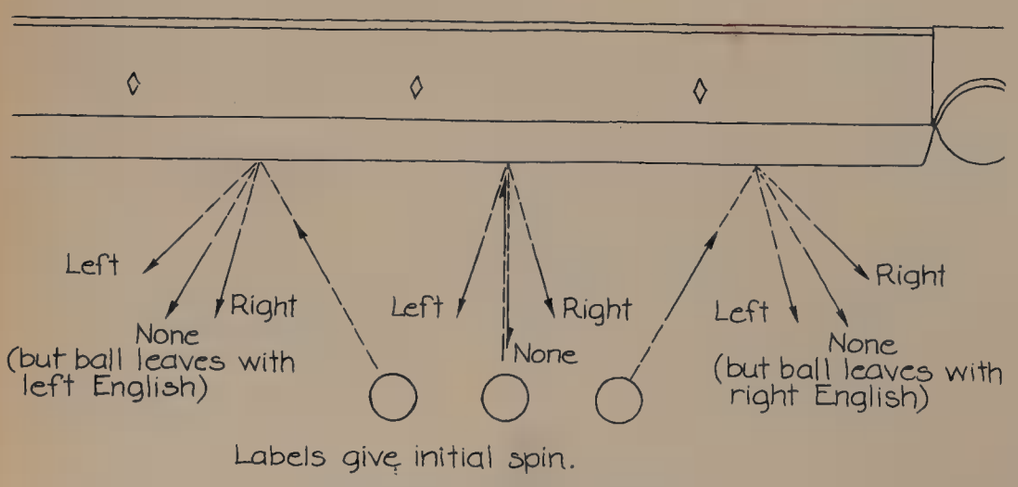
Factors in a topspin shot



Four stages in a follow shot



A follow shot demonstrated



Effects of putting English on the cue ball

off center the ball is struck, the larger the lever arm is and the faster the spin around the vertical is. The friction of the ball with the table serves only to diminish the spin.

If the cue strikes the ball high or low on the side, the ball spins about an axis that is between the horizontal and the vertical. A stroke below the middle and to the left of center results in a draw with left English. The center of mass is given momentum in the usual way and the ball spins about an axis tilted out of the vertical toward the player's left. One can view this rotation as being two simultaneous spins, one that is clockwise about the vertical and another that is backspin about the horizontal. The primary friction force on the ball is imparted by the backspin.

A cue ball with side English travels in a straight line like a ball with no English, but the angle at which it rebounds from a rail is remarkably different. When the ball has no initial English, its angle of rebound is the same as its angle of approach to the rail. If the approach to the rail is perpendicular, as is shown in the bottom illustration at the left, the ball must retrace its path after rebounding. If it has left English, however, it returns along a path on the player's left because of friction during contact with the rail. From an overhead view the ball with left English turns clockwise. When it slides along the rail, the friction force is to the left. When the ball rebounds, it has not only its velocity perpendicular to the rail but also this leftward component. The ball returns along a straight path resulting from the combination of the two motions.

A left-English ball approaching the rail at any other angle is similarly redirected. The opposite effects arise with a right-English ball. A simple way to remember the difference is to associate the direction of English with the rotation of the rebound path: left English results in a clockwise rotation of the path and right English results in a counterclockwise rotation.

If a ball approaches the rail with no English and at some angle other than 90 degrees, friction from the rail gives it English. Consider the ball's approach to be a combination of one motion perpendicular to the rail and another parallel to it. Once contact is made the parallel component generates friction on the ball. The resulting torque spins the ball about the vertical, imparting English. Suppose a ball is sent into the rail from the player's right. If it initially had no spin, it leaves the rail with left English.

All these rotations of spin are about an axis in a plane perpendicular to the direction of travel of the ball. The massé shot supplies spin about an axis out of that plane. With the cue almost vertical the player strikes downward on the side

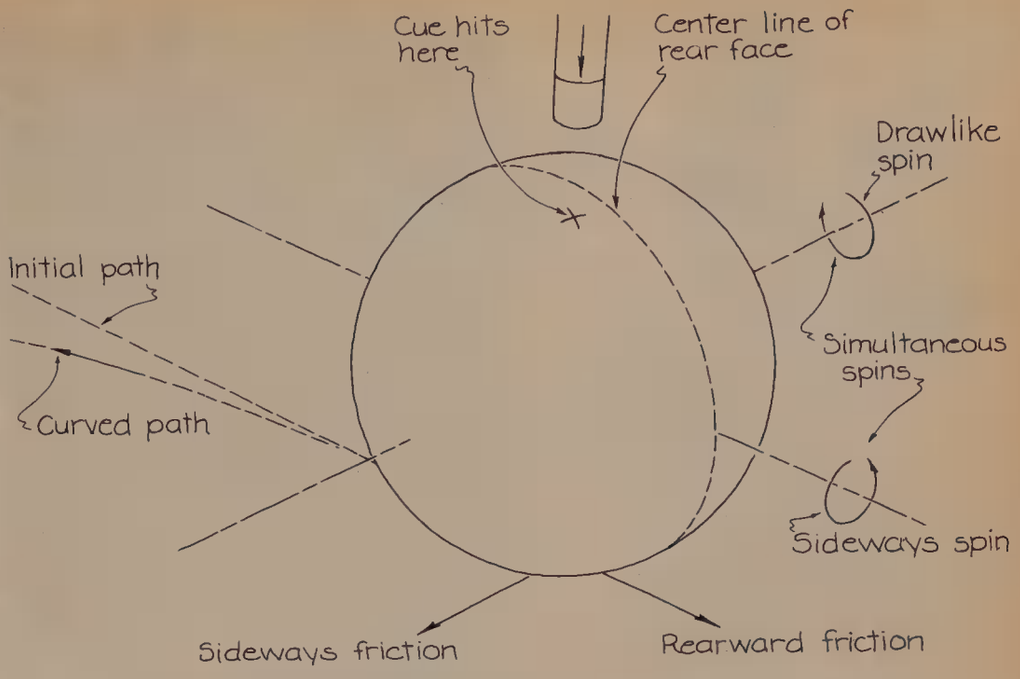
of the cue ball. The horizontal part of the stroke determines the initial path of the ball, but the spin given to the ball generates a friction from the table that ends up curving the path.

Suppose the player strikes sharply the left side of the ball. Since the blow is hard and the lever arm is large, the spin imparted to the ball is large. The spin is about an axis that is approximately in a horizontal plane but is not perpendicular to the initial path of the ball. To simplify the spin one can regard it as consisting of two simultaneous spins about different axes. One is parallel to the initial path and the other is perpendicular to it. The spin about the axis perpendicular to the initial path is similar to that of a simple draw shot. The spin about the other axis forces the ball to slip perpendicular to the initial path, thereby generating a friction force that is also perpendicular to the path. This sideways force curves the path of the ball.

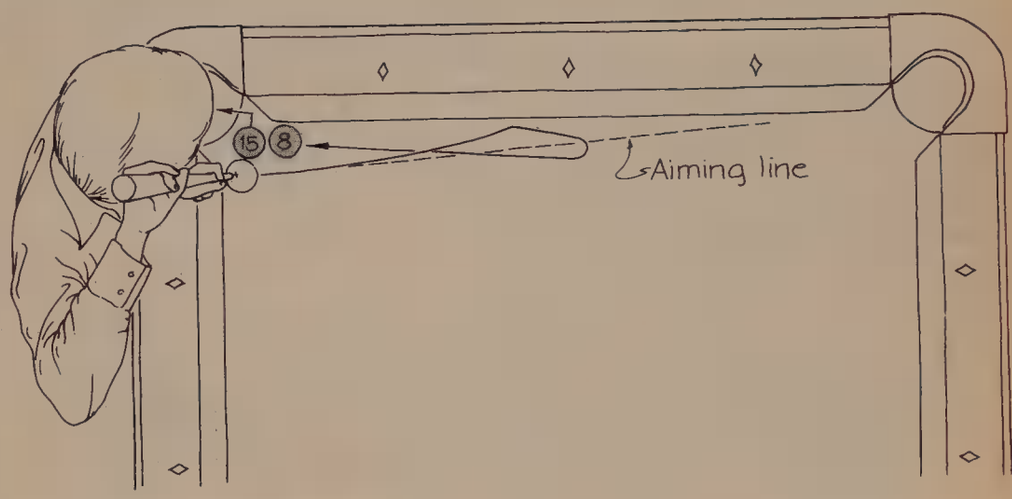
The massé is commonly employed to send the cue ball around an obstacle to reach a hidden ball. A more complex massé shot is shown in the lower illustration at the right. The idea is to sink the 15 ball and the eight ball with a single shot and to have the eight ball enter the pocket last. The cue ball is given a massé shot, knocks in the 15 ball, misses the eight ball and heads toward the rail in a curved path because of sideways friction. After rebounding from the rail the ball stops and then reverses its horizontal motion, heading back to pocket the eight ball.

The initial stroking of the cue ball gives it both backspin for a friction force like the one in the standard draw shot and sideways spin for a sideways friction to curve the path into the rail. The rebound off the rail is little affected by the backspin, but the sideways friction keeps the ball near the rail. After the rebound the backspin finally stops the horizontal motion of the ball's center of mass. Since the ball still spins, that friction force then brings it back toward the player. The remaining sideways friction continues to drive the ball toward the rail. Hence after the reversal of path resulting from the drawlike component of the shot the ball comes back near the rail to knock the eight ball into the corner pocket.

When a cue ball collides with an object ball, part of the momentum and kinetic energy of the center of mass of the cue ball is transferred. The transfers are almost total if the collision is head on. In a glancing collision the transfers cause the two balls to separate along approximately perpendicular paths. (In practice a small amount of energy is lost by the balls in the collision, and the angle between their paths is a bit less than 90 degrees. I shall disregard this complication.)



Factors in a massé shot

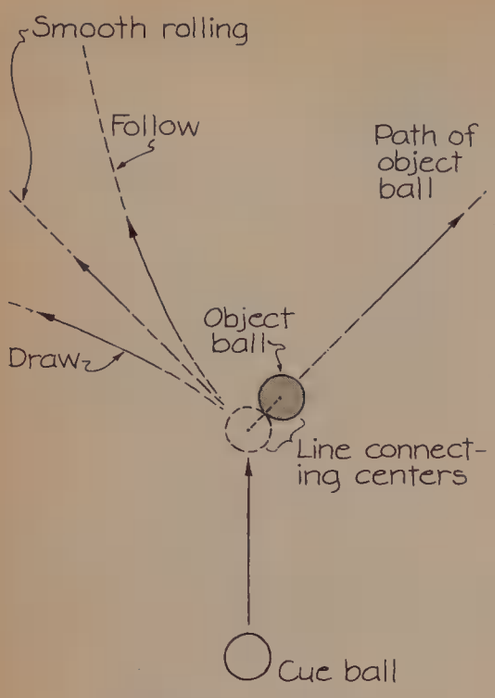


A massé shot demonstrated

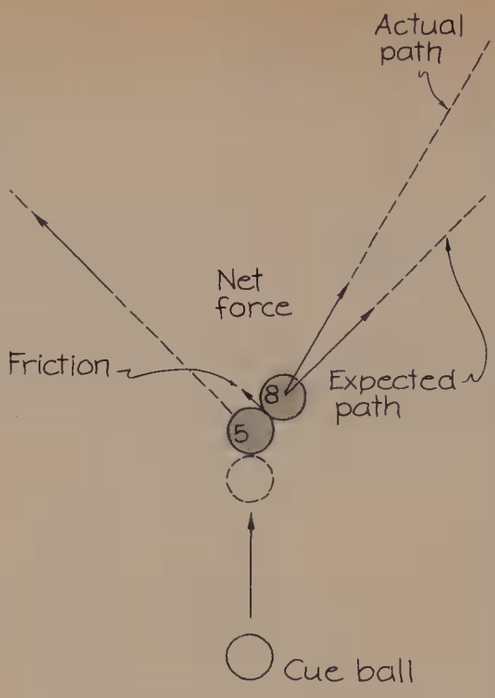
You can easily predict where the cue ball and the object ball will travel after a collision. Imagine the instant the two balls touch and mentally draw a line between their centers. At that instant the object ball acquires two forces from the cue ball. At the contact point and perpendicular to the line between the centers is a small friction force. It is almost always small enough to be ignored. Parallel to the line is a larger force that pushes the object ball off along a path that is also parallel to the line. The direction given to the object ball depends almost entirely on the orientation of the line between the centers of the balls at the instant of contact. Through experience the skilled player can direct the cue ball so that it makes contact in just the way necessary to send an object ball into a pocket. The player can be certain the cue ball will travel perpendicularly to that path.

If the cue ball is given follow or draw and still has the associated sliding when it reaches the object ball, the collision changes somewhat in that the cue ball leaves the collision site on a curved path. Suppose the cue ball has been given a large follow. The collision transfers part of the kinetic energy and momentum of the center of mass. If one can disregard entirely the friction between the colliding balls, none of the spin of the cue ball is transferred to the object ball. The cue ball begins to move away from the collision site along a path perpendicular to the path taken by the object ball. The cue ball still has topspin. The curious feature is that the ball no longer slides parallel to its path. The component of the sliding perpendicular to the path provides a sideways friction force that pushes the ball into a curved path. Therefore when the cue ball is given follow, it tends to curve back toward its original direction after a collision. A cue ball with draw tends to curve away from the original direction.

Normally the friction between two colliding balls is negligible. It can be greatly increased if the surfaces are cov-



A cue ball hits an object ball



A cue ball hits frozen object balls

ered with chalk. My favorite example comes from Byrne's book. The illustration below depicts the setup: the player must get the five ball into the pocket at the right. Can the shot be made without contact between the cue ball and the spotted ball? Normally the shot is impossible. The five ball can travel to the pocket only if the player has aligned the collision so that the line joining the centers of the cue ball and the five ball points to the pocket. The spotted ball is clearly in the way.

The shot can be made if the player chalks the left side of the five ball and then sends the cue ball into the side with a little right English. With chalk on the collision area the friction between the balls is no longer negligible. The five ball is subjected to two forces during the collision, one force parallel to the line connecting the centers and the other force (friction) perpendicular to that line. The five ball heads off in the direction of the net force, which by design is toward the corner pocket.

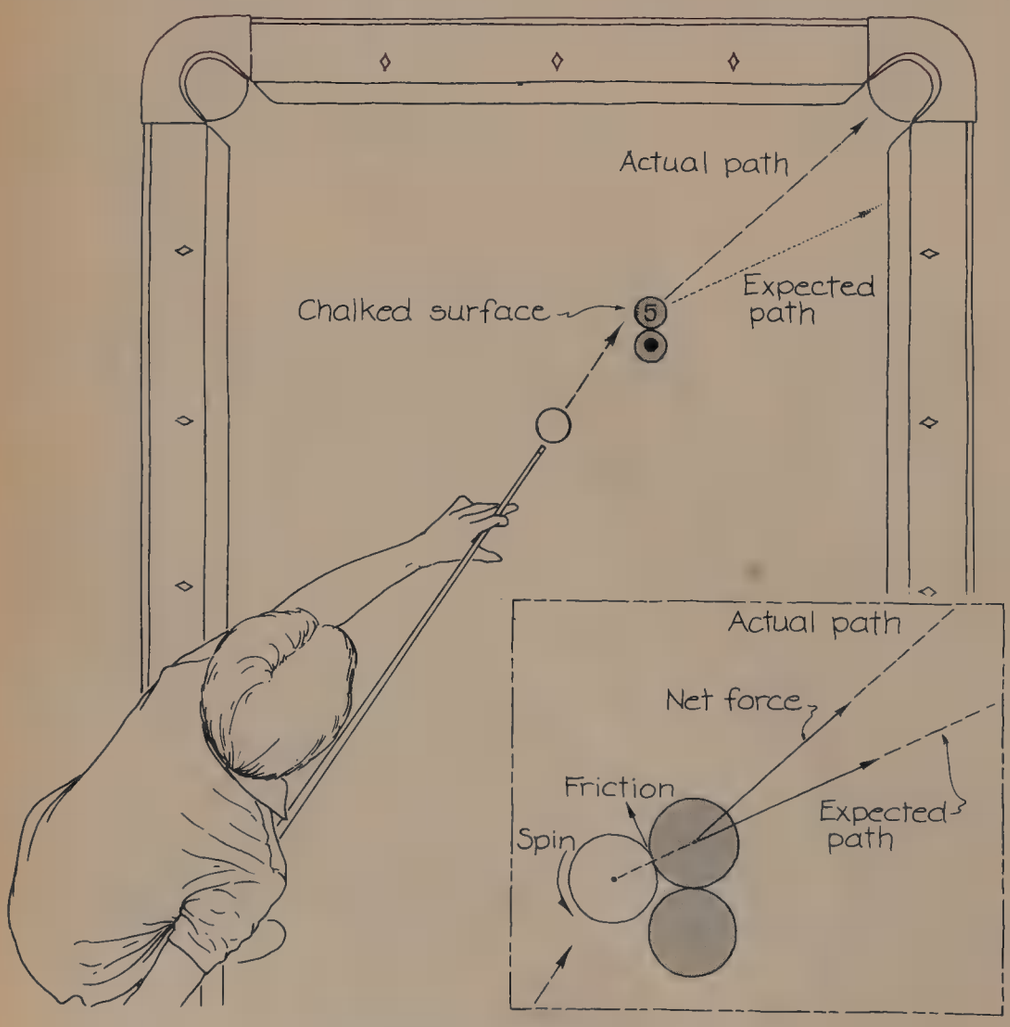
Chalking a ball is certain to get you thrown out of a pool game, but a similar application of friction between balls can be achieved more acceptably. When a cue ball strikes an object ball that is already touching another object ball (the two object balls are said to be frozen), the collision creates a friction between the object balls that can significantly alter the path of one of them. Consider the situation in the illustration at the top right on this page. The cue ball is sent directly into the five ball, which is frozen to the eight ball.

It is best to analyze the collision in two steps. First the cue ball transfers energy and momentum to the five ball, which then collides with the eight ball. The five and the eight should separate along perpendicular paths, but the eight ends up traveling more in the forward direction because of friction between the object balls.

During the collision between the five ball and the eight ball the five begins to move perpendicular to the line between the centers. The eight ball should move parallel to that line because of the force of collision from the five. Since the balls are initially frozen, however, the motion of the five ball rubs the surface of the eight ball, generating a friction force that briefly pushes the eight ball perpendicular to the anticipated path. The actual path is then set by the combination of these two forces on the eight ball during the collision; the path is more in the forward direction than it would be if the balls were not initially frozen.

My last example, the "just showin' off" shot, has become famous because Steve Mizerak, a master of pool, performed it in a television commercial. Five balls are clustered around a side pocket. The six ball lies at the mouth of a corner pocket. Can all six of the balls be pocketed with a single shot? I certainly cannot make the shot, but Mizerak is said to be successful three times out of four.

The cue ball is sent into the two ball with follow and left English. Imagine the position of the balls and the forces between them at the instant the cue ball reaches the two ball. The two ball has three forces on it. One force is parallel to the line connecting its center with the center of the cue ball. Another is along the line connecting the centers of the



Creating friction between two object balls

two ball and the three ball. Because those balls were initially frozen, the two ball also has a rightward friction force on it at the point of contact when it slides off the three ball. The net force sends the two ball over to the five ball, where it ricochets into the side pocket.

Meanwhile the three ball has been moving. When the cue ball hit the two ball, the three ball received two forces from the two ball. A force parallel to the line connecting their centers knocked the three ball hard into the rail. The second force was leftward friction generated by the rubbing of the balls when the two ball departed toward the left. (This friction arises because the two ball and the three ball were initially frozen.) The rail pushes back on the three ball, directing it straight across the table. The ball also travels somewhat to the left because of its brief friction with the two ball. The three ball ends up in the pocket on the far side.

The five ball was initially frozen to the one ball and the four ball. When the two ball hits it, the five ball is subjected to several forces. One force is along the line connecting its center and the center of the two ball. Two more forces lie along lines connecting the center of the five ball with the centers of the one and the four. In addition the five ball has friction forces from being frozen to the other balls. The net force on the five ball is neatly toward the pocket in the left corner. The net force on the one ball is toward the nearby side pocket. The net force on the four ball is toward the far left pocket. Five balls are down.

After this cluster of balls has spread the cue ball returns to the area. It had been launched with follow (for a long run) and with left English. Its collision with the two ball left it traveling toward the far rail with most of its initial spin. The spin, however, is now somewhat sideways to the path. The cue ball curves to the left, rebounding closer to the corner pocket than it would without the sideways force. Its bounce from the rail removes most of the spin. Thereafter it travels in straight lines, bouncing twice more from rails. It finally reaches the six ball at the other end of the table and pockets it.

I have of course been describing only a limited selection of shots. Thousands of interesting shots remain to be analyzed. You might be particularly interested in figuring out the physics of shots into large clusters of frozen balls. Byrne has several curious examples devised by 19th-century masters of pool and billiards. You might also be interested in jump shots, in which the cue ball is sent hopping over the table or even between two tables. Be careful. Proprietors of pool halls rarely tolerate such shenanigans, even in the interest of scientific investigation.

YOUR EXCITING NEW HOBBY!

- Enjoy fantastic savings by assembling your own organ or piano.
- It's easy. No technical knowledge required.
- Just follow our clear, pictured instructions.
- Choose from many models from portables to consoles.
- Ask about our interest-free installment plan.

WERSI
ORGAN & PIANO KITS



For Free Information Pack M41
Call Toll Free
800-233-3865
In PA 717-299-4327
or send coupon below

WERSI Dept. M41 P.O. Box 5318
Lancaster, PA 17601

☐ Free Info. Pack: ☐ Organ ☐ Piano
☐ Catalog & Demo Record *(enclose $3)*

Name _____
Address _____
City _____ State _____ Zip _____
Phone (____) _____

Rep inquiries invited



Better Than Jogging, Swimming or Cycling

NordicTrack

Jarless Total Body Cardiovascular Exerciser
Duplicates X-C Skiing For The Best Way To Fitness

NordicTrack duplicates the smooth rhythmic total body motion of XC Skiing. Recognized by health authorities as the most effective fitness building exercise available. Uniformly exercises more muscles than jogging, swimming, cycling and rowing.

Does Not cause joint or back problems as in jogging. Highly effective for weight control and muscle toning.

Easily Adjustable for arm resistance, leg resistance and body height. Smooth, quiet action. Folds compactly to require only 15 by 17 inches of storage area. Lifetime quality.

Used In thousands of homes and many major health clubs, universities, and corporate fitness centers.

Call or Write for FREE BROCHURE
Toll Free 1-800-328-5888 MN 612-448-6987
PSI 124 F Columbia Crt., Chaska, Minn. 55318

Amateur Telescope Making

Edited by Albert G. Ingalls Foreword by Harlow Shapley

BOOK ONE begins at the beginning, teaches the basics of glass grinding and how to complete the first telescope. (510 pages, 300 illustrations.)

BOOK TWO leads on into advanced methods of amateur optical work and describes new projects for the telescope maker. (650 pages, 361 illustrations.)

BOOK THREE opens up further fields of enterprise; e.g. binoculars, camera lenses, spectrographs, Schmidt optics, eyepiece design, ray tracing (made easy). (646 pages, 320 illustrations.)

SCIENTIFIC AMERICAN
ATM Dept.,
415 Madison Ave., New York, N. Y. 10017

Please send me postpaid the following AMATEUR TELESCOPE MAKING books. My remittance of $____ is enclosed:

☐ **BOOK ONE $10.00** ☐ **BOOK TWO $12.50** ☐ **BOOK THREE $12.50**

For U.S. shipments add $1.00 each; elsewhere add $2.00 each.

Name _____
Address _____
City _____ State _____

Residents of New York City please add city sales tax.
Other NYS residents please add state sales tax plus local tax

BIBLIOGRAPHY

Readers interested in further explanation of the subjects covered by the articles in this issue may find the following lists of publications helpful.

THE SLOWING OF URBANIZATION IN THE U.S.

NEW DIRECTIONS IN URBAN-RURAL MIGRATION: THE POPULATION TURNAROUND IN RURAL AMERICA. Edited by David L. Brown and John M. Wardell. Academic Press, 1980.

NONMETROPOLITAN AMERICA IN TRANSITION. Edited by Amos H. Hawley and Sara Mills Mazie. University of North Carolina Press, 1981.

POPULATION DECONCENTRATION IN THE UNITED STATES. John F. Long. U.S. Department of Commerce, U.S. Bureau of the Census, U.S. Government Printing Office, 1981.

HOW CONTINENTS BREAK UP

PROPAGATION OF AN ACCRETING PLATE BOUNDARY: A DISCUSSION OF NEW AEROMAGNETIC DATA IN THE GULF OF TADJURAH AND SOUTHERN AFAR. V. Courtillot, A. Galdeano and J. L. Le Mouël in *Earth and Planetary Science Letters,* Vol. 47, No. 1, pages 144–160; March, 1980.

PROPAGATING RIFTS ON MIDOCEAN RIDGES. Richard Hey, Frederick K. Duennebier and W. Jason Morgan in *Journal of Geophysical Research,* Vol. 85, No. 37, pages 3647–3658; July 10, 1980.

PROPAGATING RIFTS IN SEAFLOOR SPREADING PATTERNS. Tanya Atwater in *Nature,* Vol. 290, No. 5803, pages 185–186; March 19, 1981.

PROPAGATING RIFT EXPLANATION FOR THE TECTONIC EVOLUTION OF THE NORTHEAST PACIFIC: THE PSEUDOMOVIE. R. N. Hey and D. S. Wilson in *Earth and Planetary Science Letters,* Vol. 58, No. 2, pages 167–188; April, 1982.

PROPAGATING RIFTS AND CONTINENTAL BREAKUP. V. Courtillot in *Tectonics,* Vol. 1, No. 3, pages 239–250; June, 1982.

THE SALMONID FISHES AS A NATURAL LIVESTOCK

THE ORIGIN AND SPECIATION OF ONCORHYNCHUS. F. Neave in *Transactions of the Royal Society of Canada,* Vol. 3, Series 3, Section 5, pages 25–39; June, 1958.

FISH OF RARE BREEDING: SALMON AND TROUT OF THE DONALDSON STRAINS. Neal O. Hines. Smithsonian Institution Press, 1976.

SALMON RANCHING. Edited by J. E. Thorpe. Academic Press, 1980.

PROCEEDINGS OF THE SALMON AND TROUT MIGRATORY BEHAVIOR SYMPOSIUM. Edited by E. L. Brannon and E. O. Salo. School of Fisheries, University of Washington, 1981.

LEARNING IN A MARINE SNAIL

CELLULAR ANALYSIS OF A GASTROPOD (*HERMISSENDA CRASSICORNIS*) MODEL OF ASSOCIATIVE LEARNING. Daniel L. Alkon in *The Biological Bulletin,* Vol. 159, No. 3, pages 505–560; December, 1980.

CALCIUM AND CYCLIC-AMP AS SYNARCHIC MESSENGERS. Howard Rasmussen. John Wiley & Sons, Inc., 1981.

CONDITIONING: REPRESENTATION OF INVOLVED NEURAL FUNCTIONS. Edited by Charles D. Woody. Plenum Press, 1982.

PRIMARY CHANGES OF MEMBRANE CURRENTS DURING RETENTION OF ASSOCIATIVE LEARNING. Daniel L. Alkon, Izja Lederhendler and Jonathan J. Shoukimas in *Science,* Vol. 215, No. 4533, pages 693–695; February 5, 1982.

MICROELECTRONIC PACKAGING

PHYSICAL LIMITS IN DIGITAL ELECTRONICS. R. W. Keyes in *Proceedings of the IEEE,* Vol. 63, No. 5, pages 740–767; May, 1975.

MICROELECTRONICS. *Scientific American,* Vol. 237, No. 3; September, 1977.

VLSI SUPPORT TECHNOLOGIES: COMPUTER-AIDED DESIGN, TESTING, AND PACKAGING. Edited by Rex Rice. IEEE Catalog No. EHO 191–7, IEEE Computer Society Press, 1982.

IBM 3081 SYSTEM DEVELOPMENT TECHNOLOGY. *IBM Journal of Research and Development,* Vol. 26, No. 1; January, 1982.

IBM PACKAGING TECHNOLOGY. *IBM Journal of Research and Development,* Vol. 26, No. 3; May, 1982.

PAPERS ON MANUFACTURING TECHNOLOGY—PACKAGING. *IBM Journal of Research and Development,* Vol. 27, No. 1; January, 1983.

THE DEVELOPMENT OF PALM LEAVES

ANATOMY OF THE MONOCOTYLEDONS, VOL. 2: PALMAE. P. B. Tomlinson. Oxford University Press, 1961.

THE NATURAL HISTORY OF PALMS. E. J. H. Corner. University of California Press, 1966.

ANATOMY OF SEED PLANTS. Katherine Esau. John Wiley & Sons, Inc., 1977.

THE MECHANISM OF PLICATION INCEPTION IN PALM LEAVES: PROBLEM AND DEVELOPMENTAL MORPHOLOGY. Donald R. Kaplan, Nancy G. Dengler and Ronald E. Dengler in *The Canadian Journal of Botany,* Vol. 60, pages 2939–2975; 1982.

ALTERNATIVE MODES OF ORGANOGENESIS IN HIGHER PLANTS. Donald R. Kaplan in *Contemporary Problems in Plant Anatomy,* edited by William C. Dickinson and Richard A. White. Academic Press, in press.

PARTICLES WITH NAKED BEAUTY

THE UPSILON PARTICLE. Leon M. Lederman in *Scientific American,* Vol. 239, No. 4, pages 72–80; October, 1978.

OBSERVATION OF A FOURTH UPSILON STATE IN e^+e^- ANNIHILATIONS. D. Andrews et al. in *Physical Review Letters,* Vol. 45, No. 4, pages 219–221; July 28, 1980.

DECAY OF *b*-FLAVORED HADRONS TO SINGLE-MUON AND DIMUON FINAL STATES. K. Chadwick et al. in *Physical Review Letters,* Vol. 46, No. 2, pages 88–91; January 12, 1981.

EVIDENCE FOR NEW-FLAVOR PRODUCTION AT THE $\Upsilon(4S)$. C. Bebek et al. in *Physical Review Letters,* Vol. 46, No. 2, pages 84–87; January 12, 1981.

OBSERVATION OF EXCLUSIVE DECAY MODES OF *b*-FLAVORED MESONS. S. Behrends et al. in *Physical Review Letters,* Vol. 50, No. 12, pages 881–884; March 21, 1983.

CREOLE LANGUAGES

BIOLOGICAL FOUNDATIONS OF LANGUAGE. Eric H. Lenneberg. John Wiley & Sons, Inc., 1967.

RULES AND REPRESENTATIONS. Noam Chomsky. Columbia University Press, 1980.

ROOTS OF LANGUAGE. Derek Bickerton. Karoma Publishers, Inc., Ann Arbor, 1981.

ISLE DE FRANCE CREOLE. Philip Baker and Chris Corne. Karoma Publishers, Inc., Ann Arbor, 1982.

THE AMATEUR SCIENTIST

MECHANICS. Arnold Sommerfeld. Academic Press, 1964.

BYRNE'S TREASURY OF TRICK SHOTS IN POOL AND BILLIARDS. Robert Byrne. Harcourt Brace Jovanovich, Publishers, 1982.

THE DYNAMICS OF SPORTS: WHY THAT'S THE WAY THE BALL BOUNCES. David F. Griffing. Mohican Publishing Company, 1982.

ege
gan 4822

SCIENTIFIC AMERICAN



TERMITE CHEMICAL WEAPONS

$2.50

August 1983



Established 1845 **SCIENTIFIC AMERICAN** August 1983 Volume 249 Number 2

ARTICLES

28 **TRAUMA, by Donald D. Trunkey**
Accidental and intentional injuries in the U.S. take a heavier toll than cancer and heart disease.

36 **THE PURIFICATION AND MANUFACTURE OF HUMAN INTERFERONS, by Sidney Pestka** Recombinant-DNA bacteria are now producing them in large quantities.

44 **MAGNETIC FIELDS IN THE COSMOS, by E. N. Parker**
The principle of the dynamo accounts well for the magnetic fields of planets, stars and galaxies.

66 **INTERSTELLAR MATTER IN METEORITES, by Roy S. Lewis and Edward Anders**
The meteorites known as carbonaceous chondrites contain material from outside the solar system.

78 **THE CHEMICAL DEFENSES OF TERMITES, by Glenn D. Prestwich**
Termite soldiers harass intruders with sophisticated irritants, toxins, anticoagulants and glues.

88 **RATIONAL COLLECTIVE CHOICE, by Douglas H. Blair and Robert A. Pollak**
Axiomatic analysis shows that no voting system is perfect, but it can help to determine the best.

96 **THE STAVE CHURCHES OF NORWAY, by Petter Aune, Ronald L. Sack and Arne Selberg** Going back to the 10th century, they suggest that wood buildings can be permanent.

106 **DIGITAL TYPOGRAPHY, by Charles Bigelow and Donald Day**
Most type, including this type, is currently generated on the cathode-ray screen by computer.

DEPARTMENTS

6 LETTERS
50 AND 100 YEARS AGO
9 THE AUTHORS
12 MATHEMATICAL GAMES
22 BOOKS
58 SCIENCE AND THE CITIZEN
120 THE AMATEUR SCIENTIST
128 BIBLIOGRAPHY

BOARD OF EDITORS Gerard Piel (Publisher), Dennis Flanagan (Editor), Brian P. Hayes (Associate Editor), Philip Morrison (Book Editor), John M. Benditt, Peter G. Brown, Michael Feirtag, Diana Lutz, Jonathan B. Piel, John Purcell, James T. Rogers, Armand Schwab, Jr., Joseph Wisnovsky

ART DEPARTMENT Samuel L. Howard (Art Director), Steven R. Black (Assistant Art Director), Ilil Arbel, Edward Bell

PRODUCTION DEPARTMENT Richard Sasso (Production Manager), Carol Hansen and Leo J. Petruzzi (Assistants to the Production Manager), Carol Eisler (Senior Production Associate), Karen O'Connor (Assistant Production Manager), Carol Albert, Lori Mogol, Julio E. Xavier

COPY DEPARTMENT Sally Porter Jenks (Copy Chief), Debra Q. Bennett, Mary Knight, Dorothy R. Patterson

GENERAL MANAGER George S. Conn

ADVERTISING DIRECTOR C. John Kirby

CIRCULATION MANAGER William H. Yokel

SECRETARY Arlene Wright

SCIENTIFIC AMERICAN (ISSN 0036-8733), PUBLISHED MONTHLY BY SCIENTIFIC AMERICAN, INC., 415 MADISON AVENUE, NEW YORK, N.Y. 10017. COPYRIGHT © 1983 BY SCIENTIFIC AMERICAN, INC. ALL RIGHTS RESERVED. PRINTED IN THE U.S.A. NO PART OF THIS ISSUE MAY BE REPRODUCED BY ANY MECHANICAL, PHOTOGRAPHIC OR ELECTRONIC PROCESS, OR IN THE FORM OF A PHONOGRAPHIC RECORDING, NOR MAY IT BE STORED IN A RETRIEVAL SYSTEM, TRANSMITTED OR OTHERWISE COPIED FOR PUBLIC OR PRIVATE USE WITHOUT WRITTEN PERMISSION OF THE PUBLISHER. SECOND-CLASS POSTAGE PAID AT NEW YORK, N.Y., AND AT ADDITIONAL MAILING OFFICES. AUTHORIZED AS SECOND-CLASS MAIL BY THE POST OFFICE DEPARTMENT, OTTAWA, CANADA, AND FOR PAYMENT OF POSTAGE IN CASH. SUBSCRIPTION RATE: $24 PER YEAR IN THE U.S. AND ITS POSSESSIONS; $33 PER YEAR IN ALL OTHER COUNTRIES. POSTMASTER: SEND ADDRESS CHANGES TO SCIENTIFIC AMERICAN, 415 MADISON AVENUE, NEW YORK, N.Y. 10017.

SCIENTIFIC AMERICAN

CORRESPONDENCE

Offprints of more than 1,000 selected articles from earlier issues of this magazine, listed in an annual catalogue, are available at 75 cents each. Correspondence, orders and requests for the catalogue should be addressed to W. H. Freeman and Company, 4419 West 1980 South, Salt Lake City, UT 84121. Offprints adopted for classroom use may be ordered direct or through a college bookstore. Sets of 10 or more Offprints are collated by the publisher and are delivered as sets to bookstores.

Photocopying rights are hereby granted by Scientific American, Inc., to libraries and others registered with the Copyright Clearance Center (CCC) to photocopy articles in this issue of SCIENTIFIC AMERICAN for the flat fee of $2 per copy of each article or any part thereof. Such clearance does not extend to the photocopying of articles for promotion or other commercial purposes. Correspondence and payment should be addressed to Copyright Clearance Center, Inc., 21 Congress Street, Salem, MA 01970. Specify CCC Reference Number ISSN 0036–8733/83. $2.00 + 0.00.

Editorial correspondence should be addressed to The Editors, SCIENTIFIC AMERICAN, 415 Madison Avenue, New York, NY 10017. Manuscripts are submitted at the authors' risk and will not be returned unless they are accompanied by postage.

Advertising correspondence should be addressed to C. John Kirby, Advertising Director, SCIENTIFIC AMERICAN, 415 Madison Avenue, New York, NY 10017.

Subscription correspondence should be addressed to Subscription Manager, SCIENTIFIC AMERICAN, P.O. Box 5969, New York, NY 10017. The date of the last issue on your subscription is shown in the upper right-hand corner of each month's mailing label. For change of address notify us at least four weeks in advance. Please send your old address (if convenient, on a mailing label of a recent issue) as well as the new one.

Name

New Address

Street

City

State and ZIP

Old Address

Street

City

State and ZIP



THE COVER

The painting on the cover shows a group of termites attempting to immobilize an intruding ant of the large-headed species *Pheidole megacephala*. One worker termite has grasped the intruder by its right rear leg, but the main defenders of the termite nest are members of the termite soldier caste. Soldiers of this termite species, *Trinervitermes bettonianus,* are of two sizes and defend the nest by squirting an intruder with an irritating gluelike secretion (see "The Chemical Defenses of Termites," by Glenn D. Prestwich, page 78). The intruder has already been hit by several gluey squirts. Most of the small soldiers are out of glue but the soldiers' defense has entangled two of the ant's legs.

THE ILLUSTRATIONS

Cover painting by Tom Prentiss

Page	Source
12–21	Ilil Arbel
29	Donald D. Trunkey, San Francisco General Hospital
30–34	Albert E. Miller
35	Allgemeine Deutsche Automobil Club (*left*), Albert E. Miller (*right*)
36–37	Hoffmann–La Roche, Inc.
38–42	Bunji Tagawa
45	High Altitude Observatory of the University of Colorado at Boulder
46–54	George V. Kelvin
67	Glenn MacPherson, University of Chicago
68–73	Walken Graphics
74	Mitsuo Ohtsuki, University of Chicago
75–76	Walken Graphics
79	Barbara L. Thorne, Museum of Comparative Zoology, Harvard University
80	Glenn D. Prestwich, State University of New York at Stony Brook
81–85	Tom Prentiss
86	Barbara L. Thorne, Museum of Comparative Zoology, Harvard University (*top*); Manfred Kaib, University of Bayreuth (*bottom*)
87	Barbara L. Thorne, Museum of Comparative Zoology, Harvard University (*top*); Glenn D. Prestwich, State University of New York at Stony Brook (*bottom*)
88–94	Ilil Arbel
97	Ronald L. Sack, University of Idaho
98	Mittet Foto A/S
99–101	J. Dyck Fledderus
102	Ronald L. Sack, University of Idaho
103	Ronald L. Sack, University of Idaho (*top*); J. Dyck Fledderus (*bottom*)
104	J. Dyck Fledderus
107	Autologic, Inc.
108–109	Kris Holmes and Edward Bell
110	Allen Beechel (*top*); Charles Bigelow, Stanford University (*bottom*)
112–116	Allen Beechel
117	John E. Warnock, Xerox Corporation (*top*); Allen Beechel (*bottom*)
118	Allen Beechel
119	Peter Karow, URW Unternehmensberatung, Hamburg, and Donald E. Knuth, Stanford University, courtesy of *Visible Language*
123–126	Michael Goodman

SCIENCE/SCOPE

An easily processed version of a heat-resistant plastic should find new high-temperature industrial and commercial applications, as well as promote more use of advanced composites in such aerospace products as aircraft, engines, and supersonic missiles. The new Hughes Aircraft Company polyimide, which withstands temperatures of 600°F for long periods and much higher temperatures for short periods, can be processed in existing equipment. It uses a simple one-step curing process very similar to state-of-the-art epoxies. By comparison, plastics with equivalent strength and heat resistance require complicated and expensive curing procedures. The new material will be produced and marketed under the trade name Thermid® by National Starch and Chemical Corp. of Bridgewater, N.J.

Two weather satellites are being readied to monitor the Western Pacific through the end of the decade. Under contract to Nippon Electric Company of Japan, Hughes will refurbish one Geostationary Meteorological Satellite (GMS) and build another. The GMS-2 protoflight spacecraft, in storage in Japan since 1981, will be updated and renamed GMS-3a. It is scheduled for launch in August 1984, and will replace GMS-2. The new spacecraft, GMS-3b, will serve as a back-up. The satellites will provide pictures every 30 minutes, and gather other data.

NASA's space shuttle gets off the ground with support from Hughes. Astronauts train for missions on a simulator that uses a Hughes liquid-crystal projector to show what they will see outside cockpit windows. The pictures are brighter and sharper than home projection TV because the projector contains an exclusive device that draws on some of the technology used in liquid-crystal watches. Other Hughes support includes: technicians who adjust and repair flight instruments, test equipment, and ground support equipment; an instrumentation amplifier carried by chase planes; and a unique radar and communications unit that soon will serve as the space shuttle's eyes, ears, and voice.

In what may be the world's biggest aerospace cost reduction program, Hughes and its customers, including the U.S. government, have saved $1.8 billion during the past 25 years through the ideas and ingenuity of company employees. The savings were documented by the Hughes Cost Improvement Program, in which employees are encouraged to submit cost-reduction or cost-avoidance ideas on prepared forms. Last year 6,931 employees submitted ideas that saved over $250 million. One novel suggestion was to replace old vacuum pumps, used to hold semiconductor wafers in place during testing, with inexpensive fish tank pumps modified to reverse their air flow. The annual savings was $100 per pump.

Hughes Research Laboratories needs scientists for a whole spectrum of long-term sophisticated programs. Major areas of investigation include: microwave devices, submicron microelectronics, ion propulsion, lasers and electro-optical components, fiber and integrated optics, and new electronic materials. For immediate consideration, please send your resume to Professional Staffing, Dept. SE, Hughes Research Laboratories, 3011 Malibu Canyon Road, Malibu, CA 90265. Equal opportunity employer.

Creating a new world with electronics

HUGHES

HUGHES AIRCRAFT COMPANY

For more information please write:
P.O. Box 11803, Los Angeles, CA 90291

Life is an open book. If only he'd open the book.

Reading is thinking. And learning. And growing. But a kid won't read if he doesn't want to. So we're giving kids the incentive. And the books.

We're RIF (Reading is Fundamental), a national program with hundreds of local, community projects that help kids help themselves to books. Books they can choose for themselves. And keep for their own.

RIF works. But RIF only works if people work, too.

That's why we need you—or your organization—to help start a RIF program in your community. We'll help you to start going and start growing. Write RIF, Box 23444, Washington, D.C. 20024.

When it comes to getting kids to read, RIF wrote the book.

READING IS FUNDAMENTAL



This is a public service message on behalf of Reading Is Fundamental and this magazine.

LETTERS

Sirs:

Wilson G. Pond's article "Modern Pork Production" [SCIENTIFIC AMERICAN, May] ignored a nest of questions that should be of considerable interest to producers and consumers alike, that is, to the ethical issues raised by modern intensive methods of production. I shall comment on only one of these questions: our obligation not to treat the pig as a robot. As Pond notes, pigs no longer live in outdoor pens, where they root in the ground and wallow in the mud. They spend their short lives totally confined. While this may have led to some benefits for producers and consumers, namely to increased production efficiency in certain respects, and to some benefits for the pigs, namely to improvements in health during their short lives, it has also had consequences for pigs that may not be beneficial. Confinement production units do not allow behaviors such as rooting and wallowing. The impulse of a pig to engage in such behavior, however, may be strongly genetically determined. Denied the opportunity to engage in such behavior, an intelligent animal such as a pig may suffer boredom or other unpleasant psychological states....

The ignoring of ethical issues in such an article is particularly distressing. As Pond's article indicates, research to increase the efficiency of the pig in regard to feed conversion continues. If the quality of life of the pig continues to be ignored, however, we can expect that in the more efficient production units of the future the pig's life will be even bleaker than it may be today. Hopefully, future technological improvements will be made that will at least allow the pig to live a satisfactory life until it is slaughtered. This means research should be done to develop production systems that will contribute to satisfying the pig's psychological needs as well as to improving the pig's health and increasing the pig's efficiency as a converter of feed. Attention must also be given so that pigs can be transported and slaughtered in humane ways. At present it is not uncommon to break pigs' snouts for the final trip to the slaughterhouse to prevent fighting, pigs are not always rendered unconscious during slaughter, etc. Such practices cause unnecessary pain and so should be eliminated.

HUGH LEHMAN

Department of Philosophy
University of Guelph
Guelph, Ontario

Sirs:

The concerns about animal care expressed by Professor Lehman are shared by pork producers and animal scientists. The changes that have occurred in the technology of pork production, although they are partly motivated by economics, have also been motivated by concern for the welfare of the pig. The treacheries of outdoor farrowing in bad weather, of excessive losses of newborn pigs because of crushing or smothering by a 300-pound mother lying down to nurse three-pound young and of wallowing in filth and mud were overcome by the move indoors.

The pork producer of today generally is an animal protectionist and an animal welfarist in the best practical sense. Insinuations that the quality of life of the pig is being ignored reflect ignorance or misinformation. There is intense concern among producers, scientists and engineers for identifying and controlling environmental factors that may impede animal well-being. Current research in several university and government laboratories is aimed at characterizing and measuring animal stress and at continuing to develop production facilities and systems that will maximize animal comfort.

The breaking of pigs' snouts for the final trip to the slaughterhouse to prevent fighting, referred to by Professor Lehman, is, like any inhumane treatment, abhorrent to pork producers as it is to all reasonable people. Aberrant and aggressive behavior exists in all segments of society and is condoned no more by those associated with food-animal production than it is by society at large.

The human appetite for animal foods, including pork, appears to have persisted since *Homo sapiens* first appeared on the scene. Pork producers play a vital role in meeting this demand. Some have termed the efforts of the swine industry a vainglorious attempt to help feed the hungry world. Sincere, yes, vainglorious, no. Current trends point toward continued increases in pork production around the world. Pork producers are aware of their obligations as well as their opportunities. Their concerns about proper animal care will motivate them to continue to adapt new technology developed from penetrating and vigorous research currently under way in environmental physiology and animal care. Such adaptations should yield valuable benefits to consumers and the animals serving them.

WILSON G. POND

Research Leader, Nutrition
Roman L. Hruska U.S. Meat Animal Research Center
Agricultural Research Service
United States Department of Agriculture
Clay Center, Neb.

The Optimum Shape



The Optimum Shape

Researchers at the General Motors Research Laboratories have developed the first integrated system for computer design of mechanical parts with minimum mass. Optimal Shape Generation automatically optimizes the component shape in a single computer run.

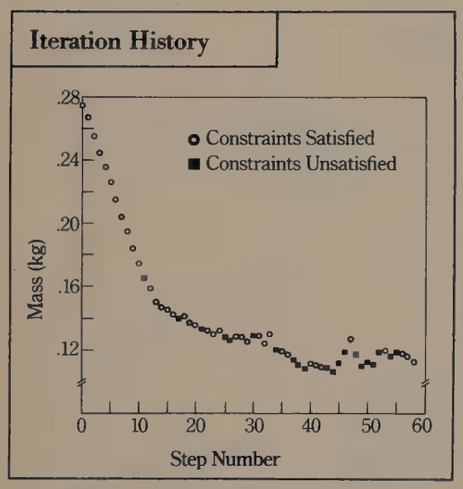


Figure 1: Decreasing mass plotted as a function of design iterations for the component shown in Figure 2.

Figure 2: Shapes as they appear on the CRT screen in the design of a minimum mass automotive component capable of performing under the structural loads. Color changes indicate (blue→yellow→green→red) increasing stress levels within the design limits.

COMPUTER-AIDED design systems automate the processes of generating geometric data and engineering drawings of parts, but they do not determine whether these parts meet structural performance requirements. In an ongoing research project at the General Motors Research Laboratories, a system has been developed that automatically ensures that the design meets structural performance constraints. More important, Optimal Shape Generation provides the component shape with the minimum mass capable of satisfying structural demands in a single computer run, without requiring human interaction with the machine.

In the last two decades, extensive research has been done in the area of computer design of structural components. Most of this work has focused on individual aspects of the process. Drs. Jim Bennett and Mark Botkin have succeeded in integrating the process from description of the model through convergence to the optimum solution.

Conventional systems continue distinctions characteristic of age-old "build and test" methods by separating the tasks of design generation and design analysis. Typically, a "designer" uses one computer system to produce engineering drawings of a given part. The task then shifts to an "evaluator" who creates a mathematical model with which to test the design on another computer system. The evaluator determines only whether or not the design meets the requirements. A lengthy interaction between the designer and the evaluator is required to optimize the design. Optimal Shape Generation integrates the process from design generation through design optimization. The system can generate the mathematical model from the design data as the shape changes without requiring additional input, thereby turning the process from a multiperson, multimachine operation into a one-person, one-machine operation.

Since there is no interaction beyond the initial input, a flexible description of the problem is crucial to effective use of the system. The researchers responded to this challenge by developing a geometric format based on a parametric description of the boundary. Defining the problem with geometric data is desirable because it describes the shape of the part in a



form directly suitable for conceptual visualization.

Because the boundary geometric description must be transformed into an analysis model not once but several times, some type of automatic finite element mesh generation is required. The researchers adapted a mesh generation technique which divides a closed region into triangular elements based on a discrete description of the boundary. The sizes of the elements of the mesh are determined by a characteristic length selected for each problem and are related to the need for accurately describing the geometry. Automatic triangulation is used to create a set of connectivities for the discrete points placed uniformly throughout the part's interior with approximately the same density as the boundary points. The combination of boundary data description and automatic mesh generation permits the system to accommodate major changes in shape from the initial design.

ADEQUACY of the triangular meshes to calculate accurate stress levels was next addressed by the development of an adaptive mesh refinement scheme. By evaluating the solution for the uniform mesh created by the choice of characteristic length and identifying areas where the strain energy density changes rapidly, the system selects the areas of the mesh that require mesh refinement. These refinements can take the form of either adding elements in the area to be refined or increasing the order of the finite element polynomial interpolation. The former approach has been taken, because it can be implemented automatically and does not require the formulation of new finite elements.

The culmination of the process introduces an optimization routine which directs the design toward a minimum mass configuration. A mathematical optimization technique is used to change the design to that shape giving minimum mass within the structural constraints. This optimization technique is based upon a sequential first-order Taylor series approximation of the constraints and a feasible directions solution of the problem. Periodic mesh refinements are performed throughout the optimization, since the design is continually changing, and the system must predict the stresses and the behavior of the constraints as the design changes.

"By taking an integrated approach," says Dr. Bennett, "we're able to combine the objectives of reducing the mass of the material and meeting structural performance requirements in a single automatic system."

"We expect," adds Dr. Botkin, "that in the future this technique will become the standard way of designing structural components."

General Motors

GM

MARK OF EXCELLENCE

THE MEN BEHIND THE WORK



Drs. Jim Bennett and Mark Botkin are members of the Engineering Mechanics Department at the General Motors Research Laboratories.

Dr. Bennett holds the title of Assistant Department Head. He attended the University of Michigan as an undergraduate and received his graduate degrees from the same institution in the field of aerospace engineering. His Ph.D. thesis concerned non-linear vibrations. Before coming to General Motors in 1973, he taught aeronautical and astronautical engineering at the University of Illinois.

Dr. Botkin is a Staff Research Engineer. He received his undergraduate and graduate degrees from the University of Missouri at Rolla. His graduate work was in the field of civil engineering, and his doctoral thesis concerned structural optimization. Prior to joining General Motors in 1978, he worked for four years as a consultant to computer applications engineers.

50 AND 100 YEARS AGO

SCIENTIFIC AMERICAN

AUGUST, 1933: "Out of years of sad experience with two kinds of destructive natural phenomena—earthquakes and hurricanes—and years of study of their forces and effects by scientists and builders, one elementary fact is beginning to emerge. This is that man has little actual need to be in terror of these angrier moods of nature if only he will provide good, substantial buildings. Good buildings will ride safely through earthquakes of major force and will outlast a heavy hurricane. It is only bad buildings that go down before these forces. We need not abandon California and Florida, and we need not even fear to move into them—provided we move into well-built structures when we get there."

"Why should sunspots be cool? A typical sunspot maintains a temperature fully 1,000 degrees centigrade below that of its surroundings, over an area thousands of miles across and for days and weeks on end. Something is obviously happening to keep it cool, and there is convincing evidence that the process of expansion in the vast ascending vortex of gases that constitutes the spot is the cooling agency. New material is continually being fed into the bottom of the vortex, deep below the surface, and as the gases rise (reversing the familiar motion of water running out of a washbasin) the forced expansion cools the newly supplied material faster than the influx of heat from the sides can warm it up. Only as the spent gases flow outward close to the surface does the heating process overtake the cooling and so set an outer boundary to the spot."

"For 88 years *Scientific American* has portrayed the scientific progress of the world. In the first decades of its existence the magazine was the sole semi-popular disseminator of such news. Latterly other magazines have sprung up, some touching only on the spectacular fringe of science. In this enlarged group *Scientific American* stands alone, the only magazine of its class that aims to present in readable style a graphic and reliable account of man's conquest of the secrets of the universe. It does not 'write down' to any group but rather aims at a high standard that will appeal to the scientist, the engineer, the professional man and the thinking layman. We cannot hope to please everyone. To those who desire information principally on mechanics or invention we recommend the other magazines that admirably report such news. Getting the proper perspective on scientific progress is *Scientific American*'s purpose. While following legitimate scientific advances rather than predicting imaginative future developments, it still will lead, as always, in accuracy and reliability."



AUGUST, 1883: "The introduction within the past two years of the improved gelatine process, by which the time of taking photographs with dry plates has been reduced a thousand times, renders it an easy matter now to obtain with certainty excellent pictures of moving objects and opens up a vast field of experiment for the scientific student. We have lately received some excellent specimens of instantaneous work by Mr. G. G. Rockwood, illustrating the principal proceedings at the opening of the great Brooklyn Bridge. Pictures of frigates covered with flying flags, sailors manning the yard arms and cannons firing from the same ships are among the photographs, and they convey to the mind an idea of the extreme brevity of the time in which the impression must have been made on the sensitive plate."

"The long voyage made in the interest of science a few years since by the *Challenger,* a ship of the British service, awakened a widespread desire upon the part of intelligent people in all portions of the world to learn something further concerning that wide and mysterious domain, the bottom of the sea. A report made by the U.S. steamship *G. S. Blake,* upon her return in February from a somewhat similar mission, has renewed this feeling. The *Blake* is a floating laboratory. She is schooner-rigged and would pass, were it not for certain accessories, for a roomy private yacht. These additions are mainly a heavy boom rigged forward and pivoting upon the foremast, and a high framework over the port bow. The former is used in handling the trawl and the latter is the complex and delicate reel and its belongings by which the miles of wire are paid out or wound in during soundings. Below decks a large proportion of the space is given over to the draughting room, laboratory and storage rooms. Every minute particle brought up during a haul is preserved and labeled. To the ordinary investigator much of this painstaking labor seems wasted, but a little reflection reveals the great utility of all the care. An example is afforded by the results of the first two years of Lt. Comdr. Charles D. Sigsbee's work upon the *Blake,* which were devoted to the Gulf of Mexico. The execution of the work embraced observations of depths, serial water temperatures and densities, and of currents when possible, together with the collection of specimens of the bottom soil and deposit, and of surface, bottom and intermedial water specimens, all of which served to add vastly to the sum of human knowledge regarding that vast and capricious basin."

"The Examiner of Interferences at the Patent Office, Mr. J. B. Church, has lately rendered a decision in the long-contested telephone case, in which the parties interested were Bell, Gray, Edison, McDonough, Dolbear, Boelker, Blake, Irwin and Richmond. Priority of invention is awarded to Bell for the art by which oral conversation or sounds of any description can be telegraphically transmitted; also for the improvement in the art of transmitting vocal sounds or spoken words telegraphically; also for the acoustic telegraph, including sound producers as well as reproducers on armature plate, the electro-magnet for the same, and a closed circuit passing from the helix of such electro-magnet to the source of undulatory electric energy; also for the telephonic transmitters and the combination in one circuit of two or more disks or diaphragms; also for the combination for rendering audible acoustic vibrations; also for the combination in an acoustic telegraph of an electro-magnet and a polarized armature, and the combination in an acoustic telegraph of an armature plate polarized by induction, a resonant tube and an electro-magnet and circuit connections. Priority of invention is awarded to Edison, although he did not claim it, for the transmitter, consisting of the combination in an electric circuit of a diaphragm and a liquid or equivalent substance of high resistance, whereby the vibrations of the diaphragm cause variations in the resistance of the electric current; also for the combination in a telegraph instrument operated by sound of two or more electrodes placed in an electrolytic liquid, and operating to increase and decrease the resistance of the electric circuit by the movement derived from the diaphragm; also for a spring forming or carrying one electrode and constantly pressing against the other electrode and the diaphragm to maintain the required initial pressure between the electrodes and yield to the movements of the diaphragm. Priority of invention for a telephone receiver, consisting of the combination in an electric circuit of a magnet and a diaphragm supported and arranged in close proximity thereto, whereby sounds thrown upon the line may be reproduced accurately as to pitch and quality, is awarded to McDonough."

THE AUTHORS

DONALD D. TRUNKEY ("Trauma") is professor of surgery at the University of California at San Francisco School of Medicine and chief of surgery at the San Francisco General Hospital Medical Center. He attended Washington State University as an undergraduate, obtaining a bachelor's degree there before going on to earn his M.D. at the University of Washington School of Medicine. In 1966 he entered the Army. After two years of service in Germany he returned to the U.S. to complete a residency in surgery at the San Francisco General Hospital. Following an additional year of training at the University of Texas Southwestern Medical School in Dallas he moved to San Francisco General in 1972. Trunkey's main scientific interest is the subject of his article; he also works on shock.

SIDNEY PESTKA ("The Purification and Manufacture of Human Interferons") is head of the laboratory of molecular genetics at the Roche Institute of Molecular Biology. Born in Poland, he came to the U.S. to receive his education. His B.A. is from Princeton University and his M.D. is from the University of Pennsylvania School of Medicine. After serving an internship in pediatrics and medicine in the Baltimore City Hospitals he joined the staff of the National Institutes of Health. He worked at the National Heart Institute and the National Cancer Institute before going to the Roche Institute in 1969. Pestka is adjunct professor of pathology at the Columbia University College of Physicians and Surgeons. He did much of the fundamental work in purifying the interferons and formulating techniques for making them in large quantities. He is an enthusiastic amateur photographer who personally photographed the first interferon crystals.

E. N. PARKER ("Magnetic Fields in the Cosmos") is Distinguished Service Professor of Physics and of Astronomy and Astrophysics at the University of Chicago. He got his B.S. in 1948 at Michigan State University. His Ph.D. in physics was awarded in 1951 by the California Institute of Technology. From 1951 to 1955 he was at the University of Utah, first as a member of the department of mathematics and then as a member of the department of physics. In 1955 he moved to Chicago. He has served as chairman of both departments of which he is currently a member. He writes: "Astrophysical observations provide a number of puzzles in classical physics with the clues to discover the previously unknown physical effects that cause them. A few of us with a skeptical disposition make our living recognizing hollow explanations [of such phenomena] and moving in to see what can be done in the way of constructing a real explanation. And this leads to the discovery of many, many hitherto unknown little effects that accumulate over the years to explain some major puzzles. The solar wind is one example; the basic structure of the galactic magnetic field is another."

ROY S. LEWIS and EDWARD ANDERS ("Interstellar Matter in Meteorites") are members of the faculty at the University of Chicago who have a common interest in the chemical aspects of astronomy. Lewis is a senior research associate. He earned his B.A. at Berkeley in 1967. He went to Chicago in 1972, shortly before his Ph.D. in atmospheric and space science was granted by the University of California at Berkeley. He writes: "My main interest is the early history of the solar system and more generally the nucleosynthetic history of the universe. My own contribution is primarily to measure the isotopic composition and abundance of the noble gases trapped in primitive meteorites." Anders is Horace B. Horton Professor of Chemistry; he has been at Chicago since 1955. A native of Latvia, he attended the University of Munich from 1946 to 1949 as an undergraduate before coming to the U.S. His M.A. (1951) and his Ph.D. in radiochemistry (1954) are from Columbia University.

GLENN D. PRESTWICH ("The Chemical Defenses of Termites") is associate professor of chemistry at the State University of New York at Stony Brook. He was graduated from the California Institute of Technology with a B.S. in 1970. His Ph.D. in chemistry, given in 1974, is from Stanford University. He served for three years as a research scientist at Cornell University and the International Centre of Insect Physiology and Ecology in Kenya before moving to Stony Brook. He is the recipient of an Alfred P. Sloan Foundation Fellowship for 1981–83 for his work on the chemistry of substances synthesized by insects. Among Prestwich' interests in relation to insects are steroid and fat metabolism and chemical ecology.

DOUGLAS H. BLAIR and ROBERT A. POLLAK ("Rational Collective Choice") are economists whose interests intersect in the subject of their article. Blair is associate professor of economics at Rutgers University. He got his B.A. at Swarthmore College (1970) before going on to obtain his

INVEST YOURSELF



A windmill to pump water for "salt farming" in India. More efficient woodburning stoves for the Sahel. Photovoltaic irrigation pumps for the Somali refugee camps.

All these are solutions to technical problems in developing countries. Devising such solutions is no simple task. To apply the most advanced results of modern science to the problems of developing areas in a form that can be adopted by the people requires the skills of the best scientists, engineers, farmers, businessmen—people whose jobs may involve creating solid state systems or farming 1000 acres, but who can also design a solar still appropriate to Mauritania or an acacia-fueled methane digester for Nicaragua.

Such are the professionals who volunteer their spare time to Volunteers in Technical Assistance (VITA), a 20 year old private, non-profit organization dedicated to helping solve development problems for people world-wide.

Four thousand VITA Volunteers from 82 countries donate their expertise and time to respond to the over 2500 inquiries received annually. Volunteers also review technical documents, assist in writing VITA's publications and bulletins, serve on technical panels, and undertake short-term consultancies.

Past volunteer responses have resulted in new designs for solar hot water heaters and grain dryers, low-cost housing, the windmill shown above and many others. Join us in the challenge of developing even more innovative technologies for the future.

VITA Putting Resources to Work for People

1815 North Lynn Street, Arlington, Virginia 22209-2079, USA

THE FOOD AND AGRICULTURAL ORGANIZATION OF THE UNITED NATIONS, Rome, Italy,

requires

AGRICULTURAL OFFICERS

The FAO/World Bank Cooperative Programme invites applications for anticipated Agricultural Officer appointments of three years initial duration, stationed in Rome, but involving frequent travel to developing countries.

The main duties involve participating in or leading missions concerned with identification and preparation of agricultural investment projects in developing countries, and being responsible for all agricultural aspects of such projects.

Essential qualifications: University degree in agriculture followed by at least 7 years of either field or managerial professional experience in agriculture under diverse conditions in more than one developing country. Practical experience of a broad range of rainfed and irrigated tropical crops. A high standard of report writing is required.

Languages: Full working knowledge of English and limited knowledge of French, or vice-versa.

Salary: Depending upon level of appointment, from US $32,000 to US $46,000 per annum net tax-free including cost of living and other allowances.

Please send detailed Curriculum Vitae *quoting this advertisement* to Mr. T.R. Short, Personnel Officer, DDC, FAO, Via delle Terme di Caracalla, 00100 Rome, Italy.



"In Just A Few Days, I'll Show You How To Do

REAL MATH

On Your Calculator!"

$$\int_a^b f \quad \sum_{n=1}^{\infty} a_n \quad \frac{df}{dx} \quad \lim_{n\to\infty}$$

•Quick. •Guaranteed.
•Easy. •Fun, Too!

INTRIGUED BY CALCULATORS? Then you can step up your math skills *fast!* Use my *new method* in guidebook form. It's called *CALCULATOR CALCULUS*. This space-travel spinoff is sure-fire, so it has a **simple guarantee** — *just return it for an immediate refund if you are not astounded at the problems you're solving with it!*

But the point is - you won't want to send it back. For this is the *easiest, fastest shortcut* ever! The day you receive your copy in the mail you'll want to put it to work. It's that exciting and helpful.

My name is Dr. George McCarty. I teach math at the University of California. I wrote this guidebook to cut through the confusion. I guide you with *examples* you follow step-by-step on your calculator — you do simple *exercises* — then you solve practical *problems* with real precision!

POWER METHODS. Need to evaluate functions, areas, volumes — solve equations — use curves, trig, polar coordinates — find limits for sequences and series? It's all here!

If you're in the biological, social or physical sciences, you'll be doing Bessel functions, carbon dating, Gompertz' growth curves, half-life, future value, marginal costs, motion, cooling, probability, pressure — and plenty more (even differential equations).

Important numerical techniques? Those algorithms are here, too: rational and Padé approximation, bracketing, continued fractions, Euler's method, Heun's method, iteration functions, Newton's method, predictor-corrector, successive substitutions, Simpson's method and synthetic division.

LOOK AT WHAT USERS SAY: Samuel C. McCluney, Jr., of Philadelphia writes: *"CALCULATOR CALCULUS IS GREAT! For ten years I have been trying to get the theory of calculus through my head, using home-study courses. It was not until I had your book that it became clear what the calculus was all about. Now I can go through the other books and see what they are trying to do. With your book and a calculator the whole idea becomes clear in a moment, and is a MOST REFRESHING EXPERIENCE. I program some of the iterative problems you suggest and it always GIVES ME A THRILL to see it start out with a wild guess and then approach the limit and stop."*

Professor John A. Ball of Harvard College (author of the book *'Algorithms for RPN Calculators'*) writes: *"I wish I had had as good a calculus course."*

Professor H. I. Freedman of the U. of Alberta, writing in *Soc. Ind. Appl. Math Review*, states: *"There can be no question as to the usefulness of this book...lots of exercises...very clearly written and makes for easy reading."*

Tektronix Engineer Bill Templeton says *"CALCULATOR CALCULUS is the best, most clearly written book I have seen for improving your math skills."*

I WANT YOU TO DO THIS. *Get my complete kit,* with a TI-35 calculator, plus its 200 p. Student Math Book, *AND* the guidebook, *ALL* for $44.95 (for shipping to USA add $2, or $5 by AIR; Foreign $5, or $10 AIR; in Calif. add $2.70 tax).

If you already have a scientific calculator, you can invest in the guidebook, *'CALCULATOR CALCULUS'*, for only U.S. $19.95 (to USA or foreign: add $1 for shipping, or $4 by AIR; in Calif. add $1.20 tax).

As pennywise Ben Franklin said, *"An investment in knowledge pays the best dividends."* GET STARTED NOW—Tax deductible for professionals.

MONEY-BACK GUARANTEE! Send for it today. Be sure to give me your complete mailing address with your check or money order. If you want to charge it (Visa or MC), tell me your card no. and exp. date. *Prompt shipment guaranteed.*

Thank you!

George McCarty

EduCALC Publications - Dept. A7
27963 Cabot Road, South Laguna, CA 92677
For fast service, phone MC or VISA orders to (714) 831-2637

TEAR OUT AND MAIL TODAY

M.A. (1972) and his Ph.D. in economics (1976) from Yale University. In 1976 he joined the faculty at the University of Pennsylvania. He left Pennsylvania in 1981 to take up his job at Rutgers. Blair's current work concerns the objectives of labor unions and the theory of bargaining. Pollak is Charles and William Day Professor of Economics and Social Sciences at Pennsylvania. He attended Amherst College as an undergraduate, getting his B.A. in 1960. His Ph.D. in economics was granted by the Massachusetts Institute of Technology in 1964. He joined the faculty at Pennsylvania the same year and has remained there since with the exception of a stint as an economist at the Bureau of Labor Statistics.

PETTER AUNE, RONALD L. SACK and ARNE SELBERG ("The Stave Churches of Norway") are structural engineers who are interested in old wood structures. Aune is senior lecturer in timber-structure design at the Norwegian Institute of Technology (N.T.H.) in Trondheim. A native of Norway, he received his training in civil engineering at N.T.H. Sack is professor of civil engineering at the University of Idaho. His degrees in civil engineering are all from the University of Minnesota; they include a B.S. (1957), an M.S. (1958) and a Ph.D. (1964). His study of the stave churches began in 1976 and 1977, which he spent on sabbatical leave as a fellow of the Royal Norwegian Council for Scientific and Industrial Research at N.T.H. Until his retirement in 1979 Selberg was head of the division of steel structures at N.T.H. He is a native of Norway who earned his degrees in civil engineering from N.T.H. His work has mainly focused on suspension bridges. He became interested in stave churches during the renovation of the church at Urnes in 1975.

CHARLES BIGELOW and DONALD DAY ("Digital Typography") combine an interest in letterforms with an interest in computers. Bigelow is assistant professor of computer science and art at Stanford University. He received his B.A. from Reed College in 1967. In 1982 he was given a MacArthur Foundation Prize Fellowship for contributions to the study of typography. With his partner, Kris Holmes, he designs typefaces for computer systems. Day is lecturer in the video department of the California College of Arts and Crafts. Like Bigelow, he attended Reed College. Since 1979 he has been collaborating with a group headed by Daniel Sandin of the University of Illinois at Chicago Circle in the design of a digital image processor. Day's interest in letterforms was aroused partly by a course in calligraphy he took at Reed; Bigelow was in the same class.

ATTEND TELECOM 83 GENEVA

Switzerland—October 26-November 1, 1983
Fourth World Telecommunication Exhibition

TELECOM 83, the fourth in a series of quadrennial World Telecommunication Exhibitions, is organized under the auspices of the International Telecommunication Union (ITU) and its 158 member countries.

Highlighting the general theme "Telecommunications for Everyone," TELECOM 83 offers a unique opportunity for individuals or groups to join in the exchange of ideas, information, and technology in the fields of telecommunications and electronics. Additionally, it will provide an excellent setting for international contacts.

An integral part of TELECOM 83 will be the Fourth World Telecommunication Forum, organized by the ITU and 50 national and international engineering societies from all five continents. Telecom and the World Telecommunication Forums are recognized as the universal and most authoritative meeting of telecommunication engineers and economists.

FORUM 83: The World Telecommunications' Summit

FORUM 83—Parts 1, 2 and 3—will assemble a "brains trust" of several thousand top executives who will present and discuss the planning, financing, management and implementation of the world telecommunication network, and the convergence of computing and communications technologies.

Over 200 presentations will enable participants to hear, meet and question industry and government leaders on current and future developments of the world telecommunication network. Plenary sessions and selected parallel sessions will have simultaneous interpretation.

Nations Working Hand in Hand

The three-day Part One of FORUM 83 will offer presentations by prominent figures in the economic sector as well as the telecommunication field. Speakers will include government leaders, senior corporate managers, chief scientists from industry, and representatives from international and financial organizations. They will discuss the technological requirements of industrialized and developing nations. They will also confront the future need for financing national, regional and international telecommunication development plans.

Towards a World Network

FORUM 83 Part Two has adopted the theme: "Integration of the World Telecommunication Network." This three-day event will be a far-reaching technical and scientific symposium with wide-ranging appeal. It represents the combined efforts of the ITU's 158 member countries to introduce to those attending, the advances made since FORUM 79 to solidify the world telecommunication network. Some 60 presentations will be made covering the latest innovations and technological trends.

Participation for each of these parts of FORUM 83 is limited to the first 1500 registrants.

In keeping with the ever-increasing transnational exchange of data, information and broadcasting, FORUM 83 Part Three, a Legal Symposium, will examine the legal aspects of international telecommunication. This concluding session of TELECOM 83 will scrutinize the international regulations relating to transnational information transport.

TELECOM 83 Exhibition

The Telecom 83 Telecommunication Exhibition encompasses all of the exhibition, conference and outdoor floorspace of Geneva's New Exhibition and Conference Centre. This quadrennial Telecommunication Exhibition will be a kaleidoscope showcasing the world's telecommunications and electronics industries, their capabilities and their networks.

The BOOK FAIR at TELECOM 83 will provide an opportunity for delegates to reference current works on telecommunications, electronics and allied fields.

Registration Fees (Swiss Francs)

	Individual	Group*	Officials**
Part 1	600	550	500
Part 2	300	275	250
Part 3	250	250	250
Parts 1 and 2	750	750	750
Parts 1 and 3	750	750	750
Parts 2 and 3	500	500	500
Parts 1, 2 and 3	850	850	850

*two or more from the same company or organization and members of institutions co-sponsoring the Technical Symposium.
**official representatives from administrations of countries/members of the ITU.

TELECOM 83 **Registration Form**

TO: **FORUM 83**
International Telecommunication Union
Place des Nations
CH-1211 Geneve 20, Switzerland

Please register me for the following Forum '83 session(s).

Part 1 ☐ at rate for individuals ☐ groups ☐ officials ☐
Part 2 ☐ at rate for individuals ☐ groups ☐ officials ☐
Part 3 ☐ at rate for individuals ☐ groups ☐ officials ☐

Name _____

Organization/Job Title _____

Address _____

_____

I enclose a cheque for SwFr of _____ **in full payment of my registration.**

I have initiated a bank transfer of SwFr _____ to Credit Suisse Geneva a/c Wagons-lits Tourisme Cooks TW 673.050.71. Specify TELECOM 83 Forum.

Date _____ Signature _____

MATHEMATICAL GAMES

Tasks you cannot help finishing no matter how hard you try to block finishing them

by Martin Gardner

EDITOR'S NOTE: Martin Gardner's column "Mathematical Games" will be appearing in this space for two issues.

> With useless endeavor,
> Forever, forever,
> Is Sisyphus rolling
> His stone up the mountain!
> —HENRY WADSWORTH LONGFELLOW,
> *The Masque of Pandora*

Suppose you have a basket containing 100 eggs and also a supply of egg cartons. Your task is to put all the eggs into the cartons. A step (or move) consists of putting one egg into a carton or taking one egg from a carton and returning it to the basket. Your procedure is this: After each two successive packings of an egg you move an egg from a carton back to the basket. Although this is clearly an inefficient way to pack the eggs, it is obvious that eventually all of them will get packed.

Now assume the basket can hold any finite number of eggs. The task is unbounded if you are allowed to start with as many eggs as you like. Once the initial number of eggs is specified, however, a finite upper bound is set on the number of steps needed to complete the job.

If the rules allow transferring any number of eggs back to the basket any time you like, the situation changes radically. There is no longer an upper bound on the steps needed to finish the job even if the basket initially holds as few as two eggs. Depending on the rules, the task of packing a finite number of eggs can be one that must end, one that cannot end or one that you can choose to make either finite or infinite in duration.

We now consider several entertaining mathematical tasks with the following characteristic. It seems intuitively true that you should be able to delay completing the task forever, when actually there is no way to avoid finishing it in a finite number of moves.

Our first example is from a paper by the philosopher-writer-logician Raymond M. Smullyan. Imagine you have an infinite supply of pool balls, each bearing a positive integer, and for every integer there is an infinite number of balls. You also have a box that contains a finite quantity of numbered balls. Your goal is to empty the box. Each step consists of removing a ball and replacing it with any finite number of balls of lower rank. The 1 balls are the only exceptions. Since no ball has a rank lower than 1, there are no replacements for a 1 ball.

It is easy to empty the box in a finite number of steps. Simply replace each ball higher than 1 with a 1 ball until only 1 balls remain, then take out the 1 balls one at a time. The rules allow you, however, to replace a ball with a rank above 1 with *any* finite number of balls of lower rank. For instance, you may remove a ball of rank 1,000 and replace it with a billion balls of rank 999, with 10 billion of rank 998, with a billion billion of rank 987 and so on. In this way the number of balls in the box may increase beyond imagining at each step. Can you not prolong the emptying of the box forever? Incredible as it may seem at first, there is no way to avoid completing the task.

Note that the number of steps needed to empty the box is unbounded in a much stronger way than it is in the egg game. Not only is there no bound on the number of eggs you begin with but also each time you remove a ball with a rank above 1 there is no bound to the number of balls you may use to replace it. To borrow a phrase from John Horton Conway, the procedure is "unboundedly unbounded." At every stage of the game, as long as the box contains a single ball other than a 1 ball, it is impossible to predict how many steps it will take to empty the box of all but 1 balls. (If all the balls are of rank 1, the box will of course empty in as many steps as there are 1 balls.) Nevertheless, no matter how clever you are in replacing balls, the box eventually must empty after a finite number of moves. Of course, we have to assume that although you need not be immortal, you will live long enough to finish the task.

Smullyan presents this surprising re-



A tree-trimming task

Radio Shack Presents
The Micro Executive Workstation

- Powerful Built-In Software
- Retains Memory Data When "Off"
- Self-Contained Telephone Modem
- 8K RAM—Expandable to 32K

$799
8K RAM
Cat. No. 26-3801

$999
24K RAM
Cat. No. 26-3802

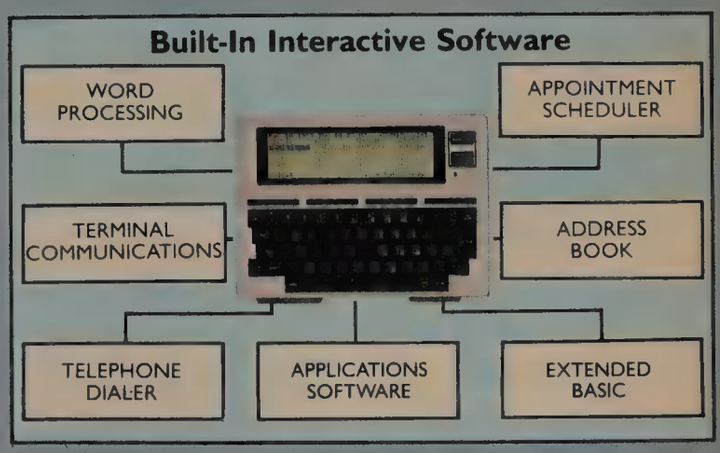


Introducing the TRS-80® Model 100 Portable Computer —User-Friendly Software Makes it Truly Revolutionary

Imagine a computer on your desk so small, it can fit in your in-basket. The second you turn it on, imagine seeing a menu of built-in executive management programs and your own files, ready for immediate use. All revealed on an eight-line by 40-character LCD display positioned just above a full-size keyboard. And when you leave the office, imagine a three-pound computer you can take along, because it works on AC or batteries.

Stop imagining. The new TRS-80 Model 100 is the computer you've been waiting for. As a desk organizer, it's a phone directory, address book, appointment calendar and telephone auto-dialer. It's a personal word processor, as well. There's even a built-in modem to access other computers by phone.

Come see the most revolutionary computer since the TRS-80 Model I at over 6500 Radio Shack stores and participating dealers, including over 400 Radio Shack Computer Centers nationwide.



Send me a free TRS-80 Model 100 brochure today!

Mail To: Radio Shack, Dept. 84-A-221
300 One Tandy Center, Fort Worth, Texas 76102

NAME _____

COMPANY _____

ADDRESS _____

CITY _____ STATE _____ ZIP _____

TELEPHONE _____

Retail prices may vary at individual stores and dealers.

Radio Shack®
The biggest name in little computers®
A DIVISION OF TANDY CORPORATION

sult in a paper, "Trees and Ball Games," in *Annals of the New York Academy of Sciences* (Vol. 321, pages 86–90; 1979). Several proofs are given, including a simple argument by induction. I cannot improve on Smullyan's phrasing:

"If all balls in the box are of rank 1, then we obviously have a losing game. Suppose the highest rank of any ball in the box is 2. Then we have at the outset a finite number of 2's and a finite number of 1's. We can't keep throwing away 1's forever, hence we must sooner or later throw out one of our 2's. Then we have one less 2 in the box (but possibly many more 1's than we started with). Again, we can't keep throwing out 1's forever, and so we must sooner or later throw out another 2. We see that after a finite number of steps we must throw away our last 2, and then we are back to the situation in which we have only 1's. We already know this to be a losing situation. This proves that the process must terminate if the highest rank present is 2. Now, what if the highest rank is 3? We can't keep throwing away just balls of rank ≤ 2 forever (we just proved that!); hence we must sooner or later throw out a 3. Then again we must sooner or later throw out another 3, and so we must eventually throw out our last 3. This then reduces the problem to the preceding case when the highest rank present is 2, which we have already solved."

Smullyan also proves that the game ends by modeling it with a tree graph. By "tree" is meant a set of line segments each of which joins two points, and in such a way that every point is connected by a unique path of segments leading to a point called the tree's root. The first step of a ball game, filling the box, is modeled by representing each ball as a point, numbered like the ball and joined by a line to the tree's root. When a ball is replaced by other balls of lower rank, its number is erased and the new balls indicated by a higher level of numbered points are joined to the spot where the ball was removed. In this way the tree grows steadily upward, its "endpoints" (points that are not the root and are attached to just one segment) always representing the balls in the box at that stage of the game.

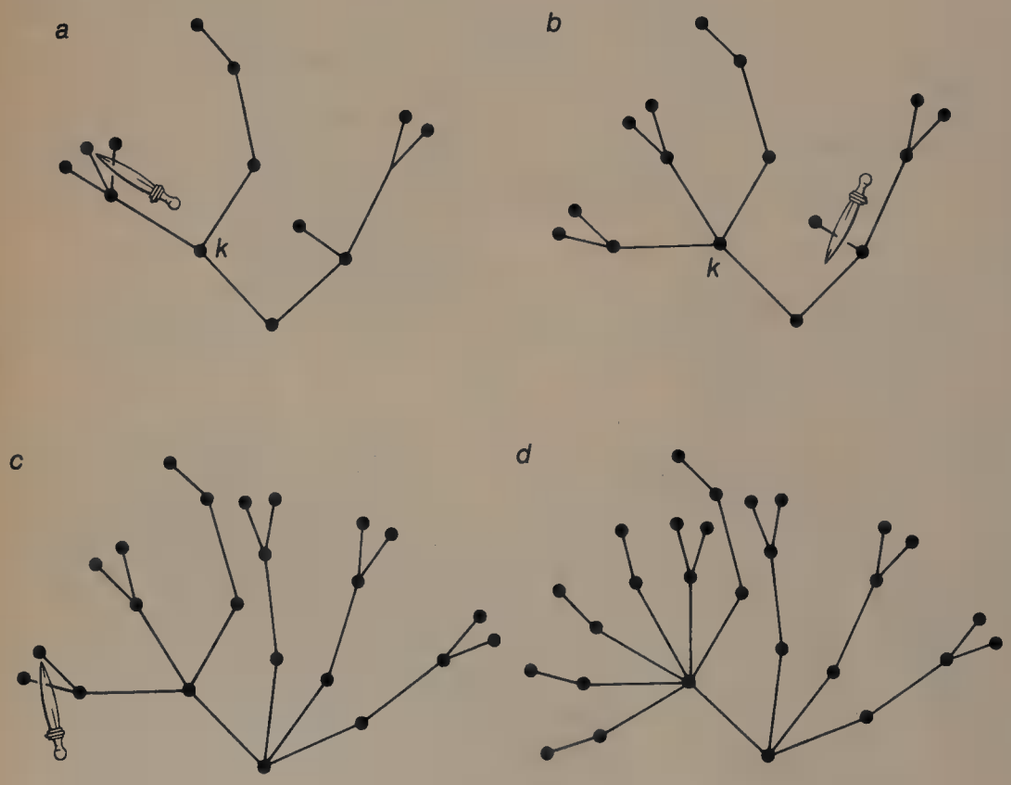
Smullyan proves that if this tree ever becomes infinite (has an infinity of points), it must have at least one infinite branch stretching upward forever. This, however, is clearly impossible because the numbers along any branch steadily decrease and therefore must eventually terminate in 1. Since the tree is finite, the game it models must end. As in the ball version, there is no way to predict how many steps are needed to complete the tree. At that stage, when the game becomes bounded, all the endpoints are labeled 1. The number of these 1 points may, of course, exceed the number of electrons in the universe, or any larger number. Nevertheless, the game is not Sisyphean. It is certain to end after a finite number of moves.

Smullyan's basic theorem, which he was the first to model as a ball game, derives from theorems involving the ordering of sets that go back to Georg Cantor's work on the transfinite ordinal numbers. It is closely related to a deep theorem about infinite sets of finite trees that was first proved by Joseph B. Kruskal and later proved in a simpler way by C. St. J. A. Nash-Williams. More recently Nachum Dershowitz and Zohar Manna have used similar arguments to show that certain computer programs, which involve "unboundedly unbounded" operations, must eventually come to a halt.

A special case of Smullyan's ball game is modeled by numbering a finite tree upward from the root as is shown in the figure at the left in the illustration on page 12. We are allowed to chop off any endpoint, along with its attached segment, then add to the tree as many new branches as we like, and wherever we like, provided all the new points are of lower rank than the one removed. For example, the figure at the right in the illustration shows a possible new growth after a 4 point has been chopped off. In spite of the fact that after each chop the tree may grow billions on billions of new branches, after a finite number of chops the tree will be chopped down. Unlike the more general ball game, we cannot remove any point we like, only the endpoints, but because each removed point is replaced by points of lower rank, Smullyan's ball theorem applies. The tree may grow inconceivably bushier after each chop, but there is a sense in which it always gets closer to the ground until eventually it vanishes.

A more complicated way of chopping down a tree was proposed by Laurie Kirby and Jeff Paris in *The Bulletin of the London Mathematical Society* (Vol. 14, Part 4, No. 49, pages 285–293; July, 1982). They call their tree graph a hydra. Its endpoints are the hydra's heads, and Hercules wants to destroy the monster by total decapitation. When a head is severed, its attached segment goes with it. Unfortunately after the first chop the hydra acquires one or more new heads by growing a new branch from a point (call it k) that is one step below the lost segment. This new branch is an exact replica of the part of the hydra that extends up from k. The figure at the top right in the illustration at the left shows the hydra after Hercules has chopped off the head indicated by the sword in the figure at the top left.

The situation for Hercules becomes increasingly desperate because when he makes his second chop, *two* replicas grow just below the severed segment [*figure at bottom left*]. And *three* replicas grow after the third chop [*figure at bottom right*], and so on. In general, n replicas sprout at each nth chop. There is no way of labeling the hydra's points to make this growth correspond to Smullyan's ball game; nevertheless, Kirby and Paris are able to show, utilizing an argument based on a remarkable number theorem found by the British logician R. L. Goodstein, that no matter what se-



Growth of the hydra



Thunderbird Turbo Coupe. A World Class Touring Car.

First impressions. As you enter Turbo Coupe you notice how the doors curve into the roof to smooth and quiet the airflow. You sit in an articulated seat. It's firm. Purposeful. Under and side thigh supports adjust to fit your body. The back angle and seat position adjust to align you with the controls. The pneumatic lumbar support adjusts to the curve of your back. Your feet feel how the pedals are positioned for heel-and-toe-operation. You test the five-speed shifter. The throws are short. Precise.

Start the engine. As it warms, consider the specifications. Four cylinders with aluminum pistons. 2.3 liters. Turbocharged, fuel-injected and monitored by Ford's EEC-IV Onboard Engine Control Computer. An engine so efficient it's rated at 145 horsepower (at 4,600 RPM per SAE standard J-1349) and is also rated at 33 estimated highway, 22 EPA estimated miles per gallon.* A triumph of technology over brute force.

Engage first gear... At this point, you will feel how the Goodyear Eagle performance tires, modified struts, four shock rear suspension and stabilizer bars work on the road. But nothing will better describe that feeling than a test drive. For more information, call 1-800-772-2100.

*For comparison. Your mileage may differ depending on speed, distance and weather. Actual highway mileage lower.

Get it together — Buckle up.

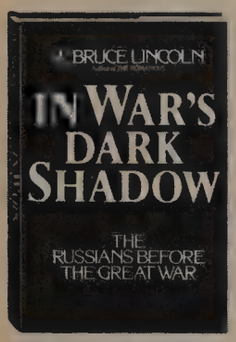
Have you driven a Ford... lately?

Ford

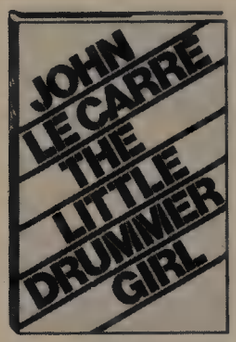
Either.

Choose any 4 books below for two dollars as your introduction to membership in the Book-of-the-Month Club.

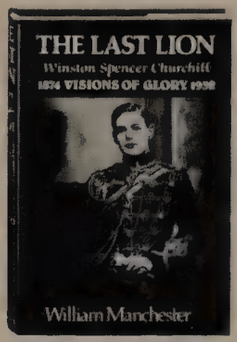
You simply agree to buy 4 books within the next two years.



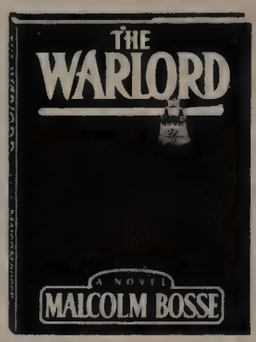
172 Pub price **$19.95**



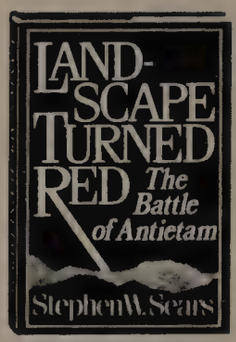
307 Pub price **$15.95**



620 Pub price **$25**



067 Pub price **$17.95**

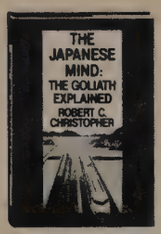


213 Pub price **$17.95**



(Software)

829 Pub price **$14.95**



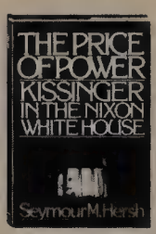
159 Pub price **$16.95**



241 Pub price **$16.95**



244 Pub price **$17.95**



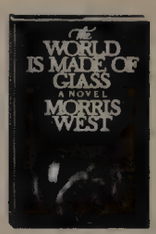
238 Pub price **$19.95**



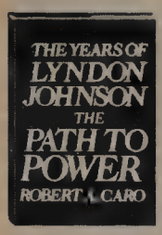
232 Pub price **$17.95**



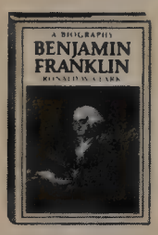
860 Pub price **$19.95**



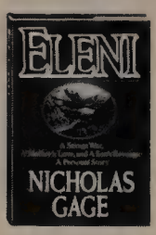
208 Pub price **$15.95**



691 Pub price **$19.95**



099 Pub price **$22.95**



593 Pub price **$15.95**



168 Pub price **$19.95**



526 Pub price **$22.95**



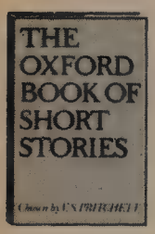
660 Pub price **$14.95**



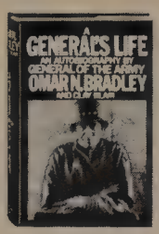
127 Pub price **$17.50**



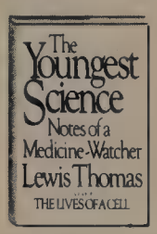
092 Pub price **$14.95**



242 Pub price **$19.95**



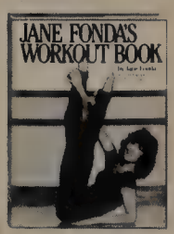
110 Pub price **$19.95**



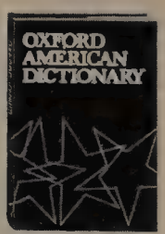
854 Pub price **$14.75**



779 Pub price **$14.95**



055 Pub price **$19.95**



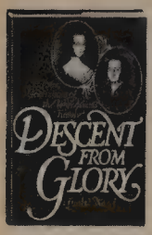
624 Pub price **$14.95**



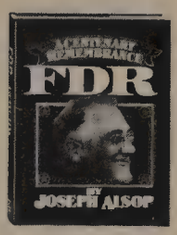
234 Pub price **$14.95**



068 Pub price **$17.95**



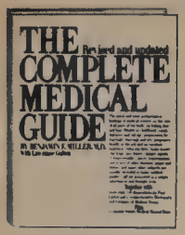
144 Pub price **$25**



786 Pub price **$25**



431 Pub price **$24.95**



722 Pub price **$24.95**



201 Pub price **$16.60**



740 Pub price **$34.95**

Benefits of Membership

Membership in the Book-of-the-Month Club begins with your choice of either 4 of today's best books for $2 or one of the extraordinary works offered here. Because our prices are generally lower than the publishers' prices, you will save on works such as these throughout your membership. In fact, the longer you remain a member, the greater your savings can be. Our Book-Dividend® plan, for which you become eligible after a brief trial enrollment, offers savings from 50% to 75% off the publishers' prices on art books, reference works, classics, books on cooking and crafts, literary sets and other contemporary works of enduring value. All Book-of-the-Month Club books are equal in quality to the publishers' originals; they are not condensed versions or cheaply made reprints.

• As a member you will receive the *Book-of-the-Month Club News*® 15 times a year (about every 3½ weeks). Every issue reviews a Selection and 150 other books that we call Alternates, which are carefully chosen by our editors.

• If you want the Selection, do nothing. It will be shipped to you automatically. If you want one or more Alternates—or no book at all—indicate your decision on the Reply Form and return it by the specified date.

• *Return Privilege*: If the *News* is delayed and you receive the Selection without having had 10 days to notify us, you may return it for credit at our expense.

• *Cancellations*: Membership may be discontinued, either by you or by the Club, at any time after you have bought 4 additional books.

• Join today. With savings and choices like these, no wonder Book-of-the-Month Club is America's Bookstore.

Or.

Choose one of the 5 valuable sets below as your introduction to membership in the Book-of-the-Month Club.

You simply agree to buy 4 books within the next two years.



The Compact Edition of The Oxford English Dictionary

for **$24.95** (Pub price $150)

The "most complete, most scholarly dictionary of the English language"—*The Christian Science Monitor.* Through photoreduction, the original 13-volume set has been reproduced in this two-volume *Compact Edition.* Magnifying glass included.



The Story of Civilization by Will and Ariel Durant

for **$29.95** (Pub prices total $335.45)

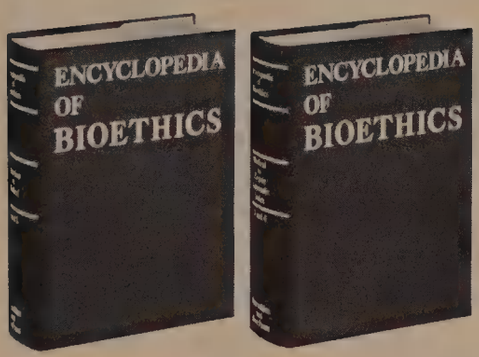
For almost half a century Will and Ariel Durant traced the continuity of world history to show the foundations of society today. The Durants' 11-volume illustrated masterwork is history come alive in all its dimensions. The enormous scope of the work covers ancient and recent civilization, including Oriental as well as Western history. A Book-of-the-Month Club exclusive for almost 50 years, it is both a reference work of undisputed authority and a lively portrait of the rulers and leaders who have shaped the course of human events.



The Encyclopedia of Philosophy

for **$24.95** (Pub price $175)

The most comprehensive encyclopedia of philosophy ever published, this set traces ancient, medieval, modern, Eastern and Western thought. An essential and rewarding reference source for home libraries.



The Encyclopedia of Bioethics

for **$19.95** (Pub price $125)

The new two-volume edition of the first comprehensive reference work in the increasingly important field of bioethics. It includes the views of 285 contributors and has 315 articles on the scientific, sociological, legal and ethical aspects of the major medical and scientific issues of today.



The Decline and Fall of the Roman Empire by Edward Gibbon Edited by J. B. Bury

for **$24.95** (Pub price $300)

The definitive Bury edition of the most acclaimed history of all. Illustrated 7-volume set, newly available on long-lasting acid-free paper, quarter-bound in leather.

Book-of-the-Month Club, Inc., Camp Hill, Pennsylvania 17012 **A170-8**

Please enroll me as a member of Book-of-the Month Club and send me *either* the four books I've listed in the "Either" boxes at right, billing me $2, plus shipping and handling charges, *or* the one set I've checked in the "Or" column, billing me for the appropriate amount, plus shipping and handling charges. In either case, I agree to buy four additional books from the Club over the next two years. A shipping and handling charge is added to each shipment.

Name_____
(Please print plainly)

Address_____ Apt._____

City_____

State_____ Zip_____

Prices shown are U.S. prices. Outside the U.S., prices are generally higher.

BOOK-OF-THE-MONTH CLUB®
America's Bookstore® since 1926.

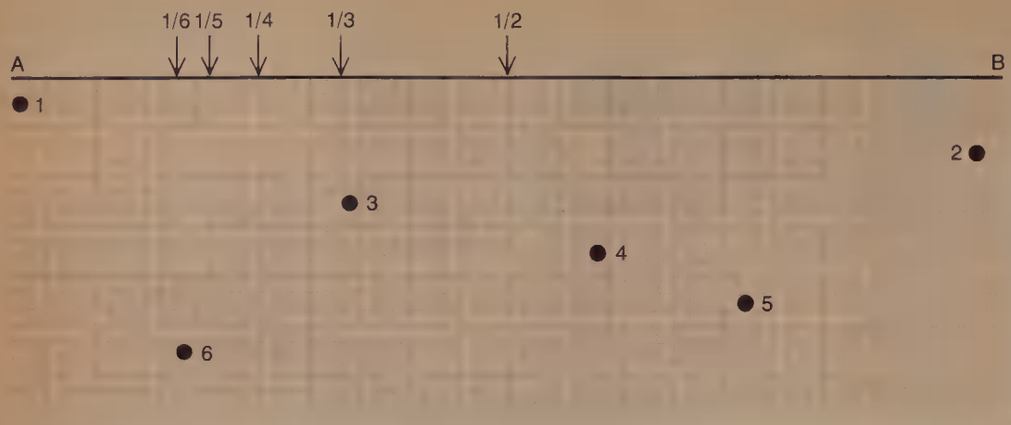
Either.
Indicate by number the four books you want

3-04

Or.
Check one box only

- 912. Compact OED **$24.95**
- 913. The Story of Civ. **$29.95**
- 917. Ency. of Philosophy **$24.95**
- 951. Decline and Fall of the Roman Empire **$24.95**
- 953. Ency. of Bioethics **$19.95**

3-64



One way to place six spots on the line A—B

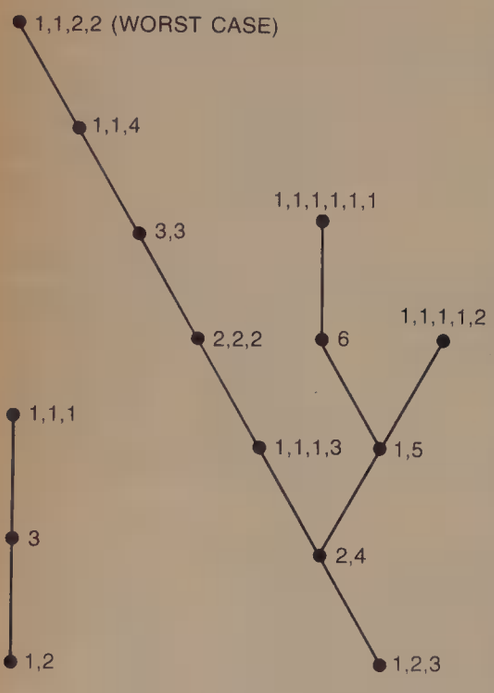
quence Hercules follows in cutting off heads, the hydra is eventually reduced to a set of heads (there may be millions of them even if the starting form of the beast is simple) that are all joined directly to the root. They are then eliminated one by one until the hydra expires from lack of heads.

A useful way to approach the hydra game is to think of the tree as modeling a set of nested boxes. Each box contains all the boxes reached by moving upward on the tree, and it is labeled with the maximum number of levels of nesting that it contains. Thus in the first figure of the hydra the root is a box of rank 4. Immediately above it on the left is a 3 box and on the right is a 2 box, and so on. All endpoints are empty boxes of rank 0. Each time a 0 box (hydra head) is removed the box immediately below gets duplicated (along with all its contents), but each of the duplicates as well as the original box now contains one less empty box. Eventually you are forced to start reducing ranks of boxes, like the ranks of balls in the ball game. An inductive argument similar to Smullyan's will show that ultimately all boxes become empty, after which they are removed one at a time.

I owe this approach to Dershowitz, who pointed out that it is not even necessary for the hydra to limit its growth to a consecutively increasing number of new branches. After each chop as many finite duplicates as you like may be allowed to sprout. It may take Hercules much longer to slay the monster, but there is no way he can permanently avoid doing so if he keeps hacking away. Note that the hydra never gets taller as it widens. Some of the more complicated growth programs considered by Dershowitz and Manna graph as trees that can grow taller as well as wider, and such trees are even harder to prove terminating.

Our next example of a task that looks as if it could go on forever when it really cannot is known as the 18-point problem. You begin with a line segment. Place a point anywhere you like on it. Now place a second point so that each of the two points is within a different half of the line segment. (The halves are taken to be "closed intervals," which means that the endpoints are not considered "inside" the interval.) Place a third point so that each of the three is in a different third of the line. At this stage it becomes clear that the first two points cannot be just anywhere. They cannot, for example, be close together in the middle of the line or close together at one end. They must be carefully placed so that when the third point is added, each will be in a different third of the line. You proceed in this way, placing every nth point so that the first n points always occupy different $1/n$th parts of the line. If you choose locations carefully, how many spots can you put on the line?

Intuitively it seems as if the number should be endless. A line segment obviously can be divided into as many equal parts as you like and each may contain a point. The catch is that the points must be serially numbered to meet the task's conditions. It turns out, astonishingly, that you cannot get beyond 17 points!



Trees for solitaire with three and six cards

Regardless of how clever you are at placing 17 points, the 18th will violate the rules and the game ends. In fact, it is not even easy to place 10 points. The illustration at the left shows one way of placing six.

This unusual problem first appeared in *One Hundred Problems in Elementary Mathematics* (problems 6 and 7) by the Polish mathematician Hugo Steinhaus. (Basic Books published a translation in 1964, and there is now a Dover soft-cover reprint.) Steinhaus gives a 14-point solution, and he states in a footnote that M. Warmus has proved 17 is the limit. The first published proof, by Elwyn R. Berlekamp and Ronald L. Graham, is in their paper "Irregularities in the Distributions of Finite Sequences," *Journal of Number Theory* (Vol. 2, No. 2, pages 152–161; May, 1970).

Warmus, a Warsaw mathematician, did not publish his shorter proof until six years later in the same journal (Vol. 8, No. 3, pages 260–263; August, 1976). He gives a 17-point solution, and he adds that there are 768 patterns for such a solution, or 1,536 if you count their reversals as being different.

Our last example of a task that ends suddenly in a counterintuitive way is one you will enjoy modeling with a deck of playing cards. Its origin is unknown, but Graham, who told me about it, says that European mathematicians call it Bulgarian solitaire for reasons he has not been able to discover. Partial sums of the series $1 + 2 + 3 + \ldots$ are known as triangular numbers because they correspond to triangular arrays such as the 10 bowling pins or the 15 pool balls. The task involves any triangular number of playing cards. The largest number you can get from a standard deck is 45, the sum of the first nine counting numbers.

Form a pile of 45 cards, then divide it into as many piles as you like, with an arbitrary number of cards in each pile. You may leave it as a single pile of 45, or cut it into two, three or more piles, cutting anywhere you want, including 44 cuts to make 45 piles of one card each. Now keep repeating the following procedure. Take one card from each pile and place all the removed cards on the table to make a new pile. The piles need not be in a row. Just put them anywhere. Repeat the procedure to form another pile, and keep doing it.

As the structure of the piles keeps changing in irregular ways it seems unlikely you will reach a state where there will be just one pile with one card, one pile with two cards, one with three and so on to one with nine cards. If you should reach this improbable state, without getting trapped in loops that keep returning the game to a previous state, the game must end, because now the state cannot change. Repeating the procedure leaves the cards in exactly



Why boomerangs in the Lands' End catalog? Mostly, just for the fun of it.

We first became intrigued with boomerangs when the U.S. Boomerang team asked us to outfit them for their historic Australian challenge tour.

We agreed because it sounded like a lot of fun. And man does not live by commerce alone.

But, along with designing their competitive clothing, we designed a line of "boomerang" sportswear for our regular Lands' End customers, too. And when the team returned from its Down Under triumphs, we invited them to model the clothing—and in the course of events, got them to give us some inside tips on throwing boomerangs as well.

The result?
Clothing plus boomerangs.

We had so much fun throwing boomerangs, we thought you might enjoy it, too. So along with the "boomerang" pants, shorts and skirts and the colorful knit shirts and sweaters that go with them, we're offering you boomerangs in our catalog, too. Authentic Australian boomerangs—in keeping with our obsession with quality in everything you get from Lands' End.

And because fun, and the clothing to wear while having it, is at the heart of the business we're in as Direct Merchants, you won't be surprised to see a colorful spread of beautiful kites in fanciful designs, along with instructions on how to fly them. And another set of pages devoted to skate sailing and Lands' End clothing suitable on the water.

In every adult, there is a child.

We take this fun-loving attitude toward our business because we sincerely believe that in every adult, of whatever age, there is a child who wants to come out and play. At Lands' End, we try to make it easy for him. And affordable.

We stress *quality* above all things. Then a *price* that represents value. And always, always *service*. The kind that offers you a free catalog to do your shopping in, by mail or by phone. By mail, the coupon below does the trick. By phone, the number is 800-356-4444. Toll-free, 24 hours a day.

Finally, a guarantee.

Finally, we offer you a guarantee so unconditional we can express it in two words—and it covers every item we sell—from clothing, to sporting goods and accessories, to our original soft luggage and all our goods from around the world.

We simply say:
GUARANTEED. PERIOD.
And we mean it.

LANDS' END
DIRECT MERCHANTS

of fine wool and cotton sweaters, Oxford button-down shirts, traditional dress clothing, snow wear, deck wear, original Lands' End soft luggage and a multitude of other quality goods from around the world.

☐ **Please send free catalog.**
Lands' End Dept. Q-21
Dodgeville, WI53533



Name _____

Address _____

City _____

State _____ Zip _____

Or call Toll-free:
800-356-4444
(Except Alaska and Hawaii call 608-935-2788)

BOOKS

Offshore oil adventures, biblical plants, the geometry of behavior, soccer madness

by Philip Morrison

OFFSHORE ADVENTURE: A PICTORIAL HISTORY OF THE NORWEGIAN PETROLEUM INDUSTRY, by Thorvald Buch Hansen, Odd Jan Lange, Håkon Lavik and Willy Håkon Olsen; color photographs by Leif Berge. Universitetsforlaget, distributed in the U.S. by Columbia University Press ($44). NEW TECHNOLOGIES IN EXPLORATION GEOPHYSICS, by H. Roice Nelson, Jr. Gulf Publishing Company, Houston, Tex. ($29.95). The long Norwegian coastline, deeply cut by the breathtaking fjords, stretches from the Skagerrak 1,000 miles north well across the Arctic Circle to Nordkapp. Offshore all the way lies a shallow continental shelf, the largest in Europe, three times the land area of Norway under a couple of hundred feet of seawater, stormy, cold but not ice-covered. Under the summer sun the sea glows blue all the long northern day; in winter gloom the storm waves can crest 30 feet high. As you read these lines there are about 3,000 people hard at work out there, heavy industrial work, around the clock.

What they do, of course, is win oil and gas. They work mostly at the southern end of the long coast, close to the median line across the North Sea where Norway's claim to the shelf abuts Britain's. Twelve hours on and twelve off they work, on great steel decks held high above the waves, roughnecks straining at cables, tool pushers at the complex consoles of the central control room. It is Norwegians who now carry out, two weeks at sea for three ashore, the oilfield tasks first developed amidst the bayous and along the Texas and Louisiana coasts. Helicopters by entire squadrons carry the crews and the mail into and out of Stavanger; the powerful workboats fetch pipe and cement and drilling mud and fresh water. Out flows the oil to the big tankers moored at the loading buoy far out of sight of any land; the gas is piped across the shallow sea floor not to Norway but to foreign landfalls at Teesside and Emden. Closer to the coast the shoreline is everywhere flanked by the deep Norwegian Trench, and a pipeline crossing that 1,000-foot depth was beyond the technology of the 1970's, although one is now being laid from new fields.

In *Offshore Adventure* clear text and strikingly beautiful photographs document the Norwegian oil industry in human, technical and institutional terms since its beginnings in 1969, when a Phillips Petroleum mobile drilling rig struck oil and gas in the Ekofisk field 250 kilometers out of Stavanger; that notable flare is shown lighting the sea. Production began on the Ekofisk in 1971. Statoil, the national oil corporation and a major operator, is now able to mark its 10th birthday, in part by sponsoring this book. The scale of the Norwegian industry is not large by world standards; some 40 percent of Norwegian oil comes to the U.S., where it slakes only a 30th part of the American thirst for imported oil. Still, that earns a tidy couple of billion dollars a year, about the value of what the Norwegians extract for themselves.

Close up, this is an industry on a dramatic scale. One photograph shows a rigger handling the block and hook of the biggest construction crane in Stavanger. The smooth form of forged steel weighs 109 tons; its great flukes engage a dozen or so loops of eight-inch steel hawsers, able to lift a hotel module of eight floors of accommodations, ready to place on the platform deck. The Statfjord B platform, the biggest to date, reaches almost 900 feet from sea bottom to crane tip. The hexagonal array of concrete storage and ballast tanks that forms its heavy base is first fitted with four towering hollow concrete columns. The entire assembly is then sunk in a fjord; it is pumped out with care to rise vertically under a 40,000-ton steel deck structure set afloat on barges. The columns lift the platform high above the level of the water for permanent service. One grand picture records the day *Queen Elizabeth II* sailed into Stavanger harbor. Beside her floated the Statfjord B steel deck, with much of the liner's look, all sheer white walls and regular openings, and across the harbor the four columns awaiting the deck towered far above the city skyline. By now those shipyards have poured seabound concrete by the megaton.

Another photograph evokes the inhospitality of the North Sea: a helicopter looks down at a layer of undifferentiated whiteness, out of which eerily feather three crane ends, a tower and a gas flame 30 feet long. Then there are the specialized workboats, powerful and wonderfully maneuverable, shipping a crosswise-mounted reversible propeller as well as the usual propulsion screws aft, able to turn on a silver krone, to judge by the wakes shown from overhead. That the local shipyards of western Norway were able to build drilling platforms competitive around the world and a line of supply vessels salable to the Russians and the Japanese is part of the Viking story: there is no drawing-board surrogate for experience with the powers of the ocean. Statoil itself has a new consulting office in Beijing, where it offers expertise to the Chinese state oil company on offshore exploration.

There have been troubles, both the tragic and the merely grave. The open sea is unforgiving even to platforms. The Bravo blowout on the Ekofisk in 1977 drew Red Adair and Boots Hansen to stem the wild flow, but not before half a tankerful of oil had stained the sea; the safety management had been inadequate. The safeguards mount. In 1980 the living platform *Alexander L. Kielland,* a converted drilling rig, had one of five steel legs torn off in a near-hurricane. A hundred and twenty-three people were lost, many while watching the evening movie; the verdict was a bad weld. The safeguards are never enough. The book gives more detail about unions, safety, finance and company structure than it does about technology, but the pictures display the engineering well. Only the diving technology, reportedly now mounting a very spirited response to the sea's cold challenge, is slighted. One also misses an index.

The third-largest seismic exploration company in the world is GECO, Geophysical Company of Norway. At this point the popular account of Norway's seaward oil makes contact with *New Technologies in Exploration Geophysics,* an insider's sharp appraisal of the turbulent state of the art of seeking invisible oil underground. Published first as a set of lively and opinionated journal articles, the book is ephemeral, but it gives the reader a quick look at an ebullient segment of contemporary high technology.

Oil is worth finding. Finding it depends on many remarkable techniques of geophysics. Yet 95 percent of all expenditures in this wealthy industry on geophysics are made in the area of reflection seismology. The sources of the seismic signal are classically small explosions, now augmented with air bubbles underwater and with vibrating plates on land, four heavy trucks shaking in synchronized FM. A hundred geo-

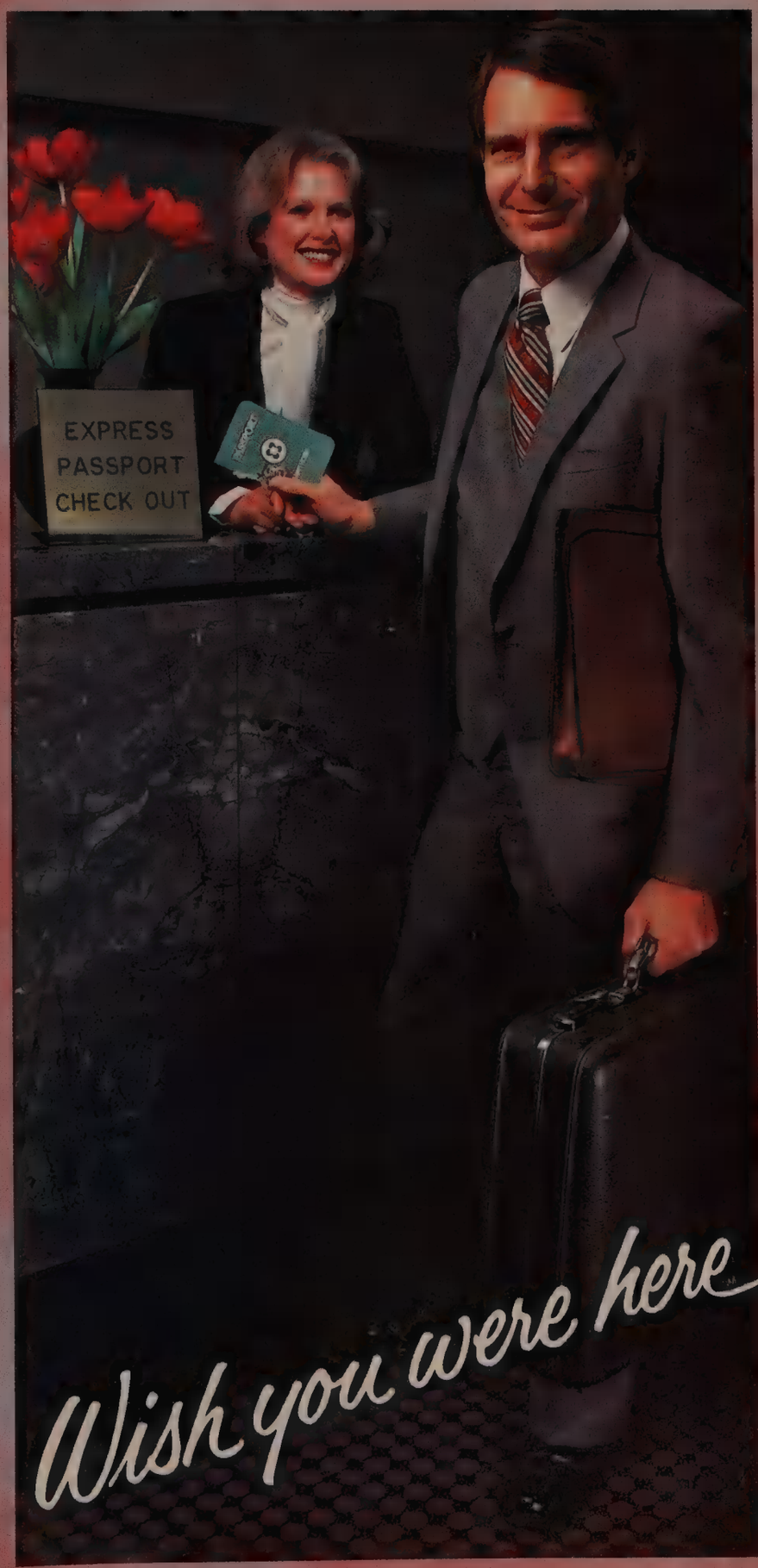
Why we make checking out of a Hyatt just as pleasant as staying at one.

The way we figure it, we give you so many extra touches at Hyatt, it would be a pity to have a long check-out line spoil them for you.

That's why when you check out of a Hyatt you simply hand in the passport you received when you checked in. You're checked out. Your bill is mailed to you within 24 hours, so there's absolutely no waiting in line. Absolutely no way you can forget how pleasant your stay was.

Any hotel can give you the essentials, but it's those extra touches ... like express check-out, that make the trip.

For reservations at 109 hotels worldwide, call your travel planner or 800 228 9000.



Chicago
In Illinois Center overlooking Michigan Avenue and Lake Michigan.

Kansas City
At Crown Center, Hyatt has health club, tennis courts, pool.

Minneapolis
Hyatt is downtown on Nicollet Mall just 2 blocks from corporate business district.

New York
Grand Hyatt on fashionable Park Avenue at Grand Central.

Washington, D.C.
Walk to Capitol Hill, government agencies, art galleries and museums.

HYATT® HOTELS

©1983 Hyatt Hotels Corp.

phones are set out to record every costly pulse. A crew will have on the frugal end a dozen or two people at work, and 100 or more where labor costs are low and many receivers are worthwhile. Where once all was cables from the dynamic mikes, nowadays the long dispersive trains of waves are collected by radio telemetry. Meanwhile channel multiplexing grows apace, and the bit stream becomes a digitized flood. The processing of three-dimensional seismic volumes, often in a hurry to learn where the likely blocks lie before someone else does, may need to handle a gigabyte per day per complex crew. Naturally the computer becomes a bottleneck. These firms are good customers for the vector-array processors feeding the supercomputers of Control Data and Cray. The aim is an interpretation of all those endless wiggles to map a physical volume with recognizable strata, all their folds in proper place.

One of the many photographs here shows the multilayer model of loaded silicone rubbers built to scale for the Statfjord production area. Once finished, the variable-density sculpture in the tank yields admirable scans to sonic probes under computer control. The interpreters learn fast. The next step is interactive interpretation at the terminal; new ingenuities of 3-D presentation are coming on stream. Right now there is a scheme of building a virtual volume of images by a vibrating mirror synchronized with the high-resolution video output from the computer. The color pictures reproduced here are impressive, even though they lack the 3-D impact.

All these data have to be stored somewhere, and the big roll stuffed on the laboratory shelf with those old surveys from the Falklands, say, seems a poor match to the FM shakers, the telemetry and the Cray. Still the reader is not quite prepared for the complaint that the footprint of the INMARSAT data-transmission satellite system misses Oklahoma City, Denver and Calgary, "one of the problems that have kept satellite data transmission out of the oil patch." One gathers that the satellite link is on the way. It would be hard to suggest any other book that so clearly paints the



image of an economic technology in such a full flood. Knowledge is bits, and enough bits is new oil. The last chapter suggests that the university supply of the trained people to examine all those costly, swift 3-D images calls for close attention and generous money inputs, not high corporate bids to lure away the best faculty. Is there a plan?

PLANTS OF THE BIBLE: A COMPLETE HANDBOOK TO ALL THE PLANTS, by Michael Zohary. Cambridge University Press ($16.95). Three little golden piles lie side by side in the photograph. They are the ripe grains of barley, emmer wheat and einkorn wheat, still grown in some fields in the Holy Land as they have been for 80 centuries. "Unlike the modern cereals, these grains still bear their shaft after threshing." Here too are ears of the wild wheat, frustrating to the would-be domesticator because the heads tend to shatter and disarticulate into single spikelets that fall to the ground before the harvest. The epic search for a mutant form that would not shatter became the domestication of wheat; those tough little bits of shaft were a small price to pay for the wholeness of the ear.

The ethnobotany in Scripture is plainly matchless for its literary quality and influence. It offers a rich view of one agricultural world, at a time about 3,000 years back and a place near the very center of origin of the crops and practices that still nourish our lives. Professor Zohary is a distinguished lifelong student of the flora of the Middle East; his own photographs add beauty and conviction to his careful plant-by-plant account of 128 species from the biblical green world. The Bible—Old Testament and New Testament together—names only 110 species; the others arise from the recognition that entire groups of plants, for instance the thorny and thistly ones, are often implied by a single term.

Plants enter the texts in everyday ways, in accounts of ritual and often unforgettably as figures of speech. Figs are still not gathered from thistles; plainly not every Hebrew or Greek usage in the long texts can be unambiguously matched to a plant species, even call-



ing on the ancient translations into the Greek, the Latin, the Aramaic and the Syriac, works much older than the European renderings and less subject to the natural error of naming biblical plants by European analogues. Botanical expertise is the major tool of the task, and to its aid the learned author, a botanist long steeped in the voluminous philological record and its living counterparts, brings the cognates, particularly in Arabic, still in use on the land.

There is no balm in Gilead, and there never was. The pungent balsam resin comes from the small fruits of a thorny desert shrub. The spade has yielded up the equipment for commercial extraction of the balm within the ruins of En Gedi on the Dead Sea. Judea was famed in antiquity for its balsam. We can forgive the scholars of King James's time, who ought to have written the name of a larger gum tree, the storax. The storax too is not found now in Gilead (the central region of the high plateau across the Jordan), but it seems probable that it grew there once, as it grows and is slashed for a sticky exudate in highland Anatolia today.

The tares sown among the wheat "while men were sleeping" was not mere parable; the grass darnel is shown in a clear photograph. Its grains are so much like those of wheat in size and shape that they cannot be removed by sieve or winnowing fan, and they may embitter the ground meal. Like rye and oats, Syrian scabious, a second form of tare, has become a crop plant on its own, although an unwanted one. It belongs to the teasel family, but long centuries of unintended selection have formed it into a plant whose seeds are indistinguishable from those of wheat. Not a stealthy enemy but the husbandman himself has unwittingly sown, reaped, threshed and sieved the tares; sometimes now he can take only a bitter harvest, the unwanted crop overwhelming the desired one.

This delightful book is handsome, inexpensive and compact. It is an ornament among the ponderous old volumes on its much-gleaned topic. In addition to the main plant list, with a page and a photograph or two for each species, it provides an admirable apparatus of indexes, maps and an introductory textual history, geography and ecology. Not every question is answered; one would like more than one can have. The forbidden fruit of the tree that was in the midst of the garden is simply not identified in the text. It may have been an apple, as widespread tradition tells us, since apples have been grown there, "eastward in Eden," toward Turkey and Syria, since about 4000 B.C., but the right word does not stand on the page. Manna is still a problem too. A sweet white exudate of the hammada shrub, a plant widespread in the southern Sinai, is shown here. A decade ago the Bedouins were still collecting those hardened droplets to use in cakes. But quantitative calculation fails; our author concludes that all the possible sources of exudate together could not provide "much more than a tidbit for the hungry people." It is no wonder that the word *man-ha* means "What is that?"

DYNAMICS—THE GEOMETRY OF BEHAVIOR: PART 1, PERIODIC BEHAVIOR, by Ralph H. Abraham and Christopher D. Shaw. Aerial Press, Inc. P.O. Box 1360, Santa Cruz, Calif. 95061 ($29). This book is an aspiring experiment in pedagogy, a systematic and extended introduction to an important branch of mathematics, one that promises a wide range of applications in addition to its powerful results of principle. It is called dynamics, since it grew from the problems and the methods of classical mechanics, turned much more general, undaunted by the nonlinearity of the real world. Its aim is to model a wide variety of systems by geometric means, seeking not the transient details of change but the overall outcome: "prediction forever" for systems from biological clocks to interacting species, parts of the world not to be grasped by the simpler formalism of the past.

The experiment is the presentation of this rich discipline at an ambitious if beginning level, almost entirely visually. The brief verbal explanations are in the form of captions to a long and steadily growing series of uniformly presented and color-coded line diagrams. The abstraction gains strength from the generality and growth of the sequence as a whole; it is revealed by pages drawn in a way recalling the pseudo realism of the thought experiments dear to Niels Bohr. The oscillators are boxes replete with cables and meters, the pendulum swings by a sturdy bearing, and so on. Capsule biographies with images of the great pioneers from Galileo to Lord Rayleigh and Balthasar van der Pol open the book, and an occasional rock group or electrocardiogram subject appears in an enlivening caricature. Nearly all formulas are suppressed in the text, to appear safely caged in a tight appendix.

About a third of the way into the volume we see a familiar system, the simple pendulum. The representation has opened up the to-and-fro and the encircling swings into a flow mapped in the phase plane, rotation rate against angle. Friction is no difficulty for these geometric methods; the central point is revealed as an attractor, toward which every nearby trajectory in phase will sooner or later come. By two-thirds of the way the phase plane has grown to three dimensions, representing a pendulum and the phase of its forcing oscillator, and even to four, although now a second reduction occurs, and a two-dimensional phase diagram for coupled oscillators is open to study on the page. The discourse is rich enough to contain the phenomena of entrainment, with its braided orbits, repellers caught between attractors as they wind. A fuller account draws out the entire invariant torus in the three-dimensional phase space, we see basins and their separatrices and we arrive at the edge of classical dynamics, as it stood in about 1950. Ahead are two more volumes of this Visual Mathematics Library, one on stable and chaotic systems and the other on the bifurcation theory and its classifications of dramatic change.

The experiment is not to be evaluated yet; the authors are persuaded of its importance in our fragmented culture, and eager for the success of their nonmathematical readers. They do, however, have very high standards for the visual grasp of complicated meaning. Ralph Abraham is professor of mathematics at the University of California at Santa Cruz, a chief contributor to the growth of modern dynamics, and Christopher Shaw is an artist and illustrator of diverse training and interest in science. Their work is a challenge above all to the small industry of textbook authors and instructors in beginning college physics and mathematics everywhere. In a digitally powerful world, in a world that needs and builds models for systems far from that lamp swinging gently in Pisa, ought we not to begin to enrich the exact solutions of simple systems, all those sines and cosines and linearized frictionless models, with the masterful if approximate geometries of our own day? This important book sets out one path to that future.

SOCCER MADNESS, by Janet Lever. University of Chicago Press ($17.50). The enormous ring is the Maracanã Stadium in Rio de Janeiro, largest in the world; it can hold 220,000 fanatic enthusiasts, separated from the playing field—and the referee—by a 10-foot moat. It is the flower of Brazil's garland of 4,000 stadiums, with a seating capacity overall about a third more than all the "civic and private stadia, bowls and ballparks" of the U.S., even though Brazil has about half our population. Brazil is first in the world for stadiums and for soccer.

Dr. Lever is a reflective sociologist whose functional study of this phenomenon arose by chance but is now solidly set in a decade of inquiry in the field. She has interviewed everyone from the peerless Pelé to the earnest fans, including a sample of 200 working-class men in Rio. Though her sample was not a properly random one, she cagily took pains to see that it mirrored the city census for age, marital status, religion, race and income

within that class. Virtually every male followed professional soccer. Only one of the men in the sample had not watched the television display of the legendary final World Cup match Brazil won in 1970, a victory followed by a welcome back to Rio that became a national holiday, brought a crowd of two million into the streets and set off a two-day carnival ecstasy that saw 44 dead and 1,800 injured. A fourth of the men in the sample were strong fans who listen to every soccer game they can, go to at least a couple of games a month and score nearly 100 percent on the test items about players set in the interview. Half of the men in the sample make soccer a routine part of their life, see a game a month and miss out on some of the trickier questions.

How did it happen? The Fédération Internationale de Football Association (FIFA) has ruled the global professional sport since 1904; it binds 147 national members, second only to the United Nations. Soccer—most of the world calls it football—is clearly a handcrafted British product of high quality, like parliaments and the Rolls-Royce, that became dispersed over the world in the halcyon days of imperial Britain. The port cities of Latin America, then hosts to large British colonies, acquired the game early. Everywhere it moved from being a sport of the gentry to being a professional game appealing to the working masses; it is still the case that the volunteer managers of the teams are amateurs, in contrast to the big business of the gate receipts, the national football lottery and the television coverage. Among the top officers of FIFA only a full-time secretary general in Zurich draws a salary. The U.S. offers "an anthropological oddity": our mass team sports are mostly, if not yet entirely, businesses run for profit. In Brazil and most other countries the football club is really a club, usually a money-losing affair with a shifting set of volunteer patrons and officials and a devoted dues-paying cadre of supporters, who may benefit from athletic facilities and others open to members. A team could hardly move to another city: civil war!

It was the play of history that seems to have made football so preeminent in Europe and Latin America. The game arrived suddenly, to find nearly everywhere no indigenous team sport on the scene. Its simple rules, its open, visible, varied and uninterrupted flow and its freedom from equipment allowed it to catch on. In the U.S. and in commonwealth countries such as Australia and India, however, professional baseball or cricket were already in place. In most countries soccer was the first professional team sport and remains the only one. It is the first team sport for amateur participation as well; indeed, there is no better predictor for a Rio man's fandom than his early personal engagement in the sport. It is not only the players who are fostered on the sandlots but also the lifelong fans.

Dr. Lever sees soccer as a profound and benign paradox of social cohesion in Brazil. It ritualizes conflict, both that on the field and that in objective social reality. The strongest rivalries in Brazilian soccer are intracity ones. There are a dozen Rio soccer clubs. Flamengo is the team of the masses, its followers millions of the urban poor, even of the steep favelas. Fluminense is thought of as a club of the elite; its applicants are screened. Criminals and the handicapped are ineligible for the Fluminense club, except "those maimed while fighting for their country or while in the service of the [club]." Its own players, "worshiped on the field," are barred from most of its social activities. This dramatic and genuine conflict is played out, however, within rules. Picture a game between Flu and Fla, 100,000 fans of each filling Maracanã, giant flags, firecrackers, confetti, team shirts, intensity—and one referee.

Soccer is no mere opiate. It integrates the nation, bonding as it divides, serving both pluralism and unity. Intercity U.S. sport emphasizes integration; Brazilian intracity soccer dramatizes cultural pluralism, in a playful sublimation of sharp conflict. Such rule-bound drama does not survive times of intense social unrest; games in Northern Ireland see the warring communities kept apart in the stadium by barbed wire and riot preparations, and the 1969 football playoff between Honduras and El Salvador, at a period of deep animosity over jobs and immigration, catalyzed a month of armed combat with thousands of casualties. The sport promotes nation building, and nationalism is the most passionate and broadest loyalty of our time. Indeed, governments can and do exploit sport; the author does not blink that fact, but she judges that the game is worth the candle, that the glowing reality of voluntary organization and civil unity overweighs the shoddy propaganda ploys of a transient regime.

There is a yawning gap in the map of unity. In Brazil football is not for women. They belong in church, men only in the stadium. The sample confirms this in detail, as do the stated views of Brazilian fans at every level. There are signs of change; Pelé has spoken up for it. Sandlot games with girls taking part will mark the real change, whenever it comes. Lever has given the reader a small book as well written as it is thoughtful; the role of sport in human society is deserving of more study, and this account is a happy example painted in the bright colors and sharp contrasts of Brazilian life.

Introducing

PABSoft

MATH MASTER SERIES

Matrix Master $24.95
+, −, *, /, trn., inv., IO, etc.

Poly Master $24.95
+, −, *, /, int., diff., etc.

* Callable from BASIC
* Fast Machine Code
* Loads Routines as needed
* TRS-80 MODS I, III, IV, Apple II+, IBM PC
* Complete Math Library
* When ordering, please describe your system, (RAM, DOS, etc.)
* Send for Scientific—Engineering Catalog

PAB Software, Inc.
P.O. Box 15397
Fort Wayne, IN. 46885
(219) 485-6980

Master-Visa accepted

Trademark Credits
TRS-80: Tandy Corp.
Apple: Apple Computer Corp.
IBM: IBM Corp.

BINOCULARS



Precision Optics From
CELESTRON®

Our Giant Series Binoculars (80mm diameter!) allow you to explore the Moon and Outerspace as well as your own backyard. Celestron also offers extremely compact roof prism and classical binoculars, and superb quality Telescopes, Spotting Scopes and Telephotos. **FREE** product, price and ordering information by returning the coupon or send $3.00 for 52-page color catalog.

CALL TOLL FREE 1-800- 421-1526

CELESTRON INTERNATIONAL
P.O. Box 3578-SA, 2835 Columbia St.
Torrance, CA 90503

Name _____
Address _____
City _____ State _____ Zip _____

Established 1845 SCIENTIFIC AMERICAN August 1983 Volume 249 Number 2

Trauma

Accidental and intentional injuries account for more years of life lost in the U.S. than cancer and heart disease. Among the prescribed remedies are improved preventive efforts, speedier surgery and further research

by Donald D. Trunkey

Trauma is the medical term for a personal injury or wound. Including both accidental and intentional injuries, physical trauma is the principal cause of death among Americans between the ages of one and 38. In 1982 there were about 165,000 deaths from trauma in the U.S., and for each death there were at least two cases of permanent disability. Statistics compiled by the Department of Health and Human Services indicate that for Americans between the ages of 15 and 24 the combined death rate from motor-vehicle accidents, homicides and suicides has risen by 50 percent since 1976. Among young whites motor-vehicle accidents are the leading cause of death, accounting for about 40 percent, whereas among young blacks homicide is the leading cause of death, accounting for approximately the same percentage. In large cities black males have a one-in-20 chance of being murdered before the age of 30. Increased urban violence has been a major contributor to the rise in the national homicide rate: from 8,464 in 1960 to more than 26,000 in 1982. Overall the death rate for American teenagers and young adults is 50 percent higher than it is for their contemporaries in other industrialized societies.

Because trauma primarily affects people at or near the beginning of their most productive work years, its cost measured in lost productivity from both death and disability is high: more than $63 million per day in lost wages from accidental trauma alone, according to a recent estimate by the National Safety Council. The total annual cost of accidental trauma, including lost wages, medical expenses and indirect work losses, comes to about $50 billion.

Trauma patients currently take up a total of about 19 million hospital days per year in the U.S., more than the number needed by all heart-disease patients and four times the number needed by all cancer patients. In the past decade the death rate from heart disease and stroke has fallen by 22 and 32 percent respectively. In contrast the death rate from accidents has risen by about 1 percent per year since 1977. Trauma is clearly a major medical and social problem in the U.S. To a large extent, however, it is being neglected by physicians, hospital administrators, government officials and the general public.

Data from several parts of the country show that death from trauma has a trimodal distribution: when the death rate is plotted as a function of time after injury, three peaks appear in the resulting graph [*see illustration on page 31*]. The first peak, characterized as "immediate deaths," represents people who die very soon after an injury. Invariably these deaths are caused by lacerations of the brain, the brain stem, the spinal cord, the heart or one of the major blood vessels. Only a fraction of the patients in this category could in principle be saved, even under the most favorable medical conditions.

The second peak, characterized as "early deaths," represents people who die within the first few hours after an injury. These deaths are usually caused by major internal hemorrhages of the head, the respiratory system or the abdominal organs, or by multiple lesser injuries resulting in severe blood loss. Almost all injuries of this type are considered treatable by currently available medical procedures. The interval between injury and definitive treatment, however, is critical to the probability of recovery.

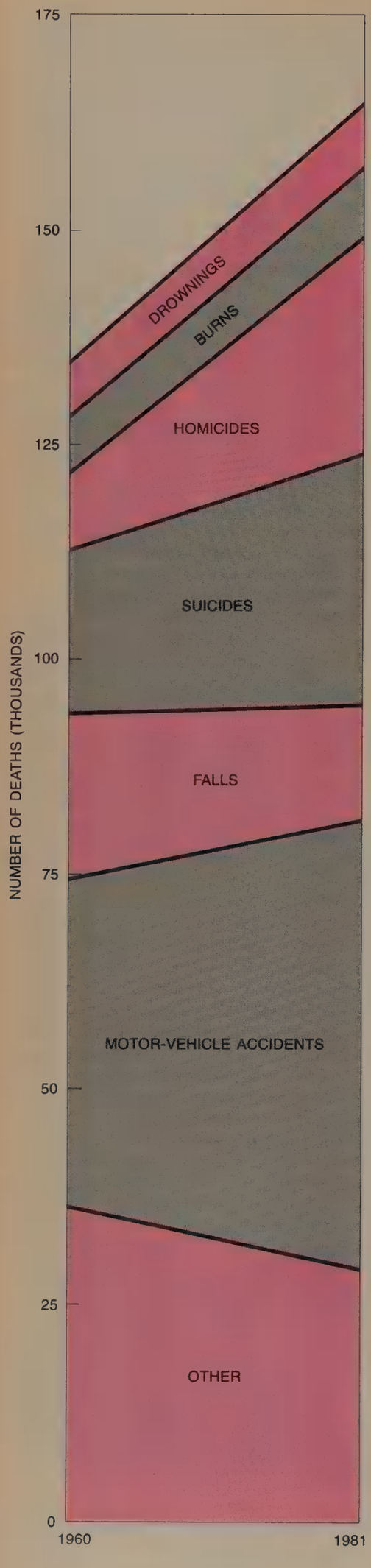
The third peak, characterized as "late deaths," represents people who die days or weeks after an injury. In almost 80 percent of these cases the cause of death is either infection or multiple organ failure. Here time is less of a factor than the quality of medical care and the extent of medical knowledge. In what follows I shall discuss the pathology of each peak in somewhat greater detail, with particular reference to the prospects for reducing the rate of mortality and disability resulting from the associated set of medical conditions.

More than half of all trauma deaths are classified as immediate. The small number of patients in this category who could be saved are those in the few large cities where rapid transportation is available and special facilities called trauma centers are in operation. A trauma center is a hospital where the medical staff has made a commitment to provide 24-hour "in house" coverage by surgeons, anesthesiologists and supporting staff to care for trauma patients. Recent medical records from two of these centers, one in Seattle and the other in San Francisco, indicate that if there are signs of life at the scene of an accident or on the way to the hospital, 20 percent of the patients who are classified as "dead on arrival" can be resuscitated in the emergency room and will eventually leave the hospital without permanent neurological damage.

This remarkable rate of recovery will probably never be achieved in most suburban and rural settings, because of the longer time it usually takes there between injury and definitive treatment. The only way to reduce the number of immediate deaths in these circumstances is through prevention. Perhaps



OPERATING ROOM at San Francisco General Hospital is the scene of efforts by surgeons, anesthesiologists and other specialists to treat a critically wounded patient. Medical records from trauma centers established in such major metropolitan hospitals indicate that the elapsed time between an injury and definitive surgical care is a critical factor in determining the survival rate of trauma victims.



TRENDS IN MORTALITY from trauma in the U.S. are plotted according to the cause of death in this graph, representing data gathered by the National Center for Health Statistics. The figures for 1960 are either actual totals for that year or averages for the period from 1952 through 1963; figures for 1981 are based on a 10 percent sample of that year's deaths.

as many as 40 percent of all deaths from trauma could be averted by the introduction of various prevention programs. Most of these programs involve controversial social issues, however, and so their chances of success are unpredictable. I shall cite here just a few of the more important trauma-prevention programs that have been proposed.

According to the Insurance Institute for Highway Safety, between 50 and 60 percent of the fatal motor-vehicle accidents in the U.S. are caused by drunk drivers. Efforts to reduce drunk driving by increasing the penalties for infractions have generally failed in the U.S., and similar programs in Europe have had only mixed results. For example, reports from a number of Scandinavian countries indicate that after mandatory jail sentences for drunk driving were imposed a significant reduction in fatal accidents was observed. In time, however, there was usually a reversion to the same mortality rate that had prevailed before the stronger measures were introduced. Rehabilitation programs for drunk drivers, introduced in several parts of the U.S., have also been found to be ineffective.

In spite of this generally negative record there is some evidence that the suspension or revocation of a driver's license after a drunk-driving conviction can have a significant effect on the subsequent rate of drunk-driving arrests in the affected population. Recently a "grass roots" group called Mothers against Drunk Drivers (MADD) was organized in California to promote such stronger measures to reduce the carnage caused by drunk drivers. The impact of this campaign, which is now spreading to other parts of the country, remains to be seen.

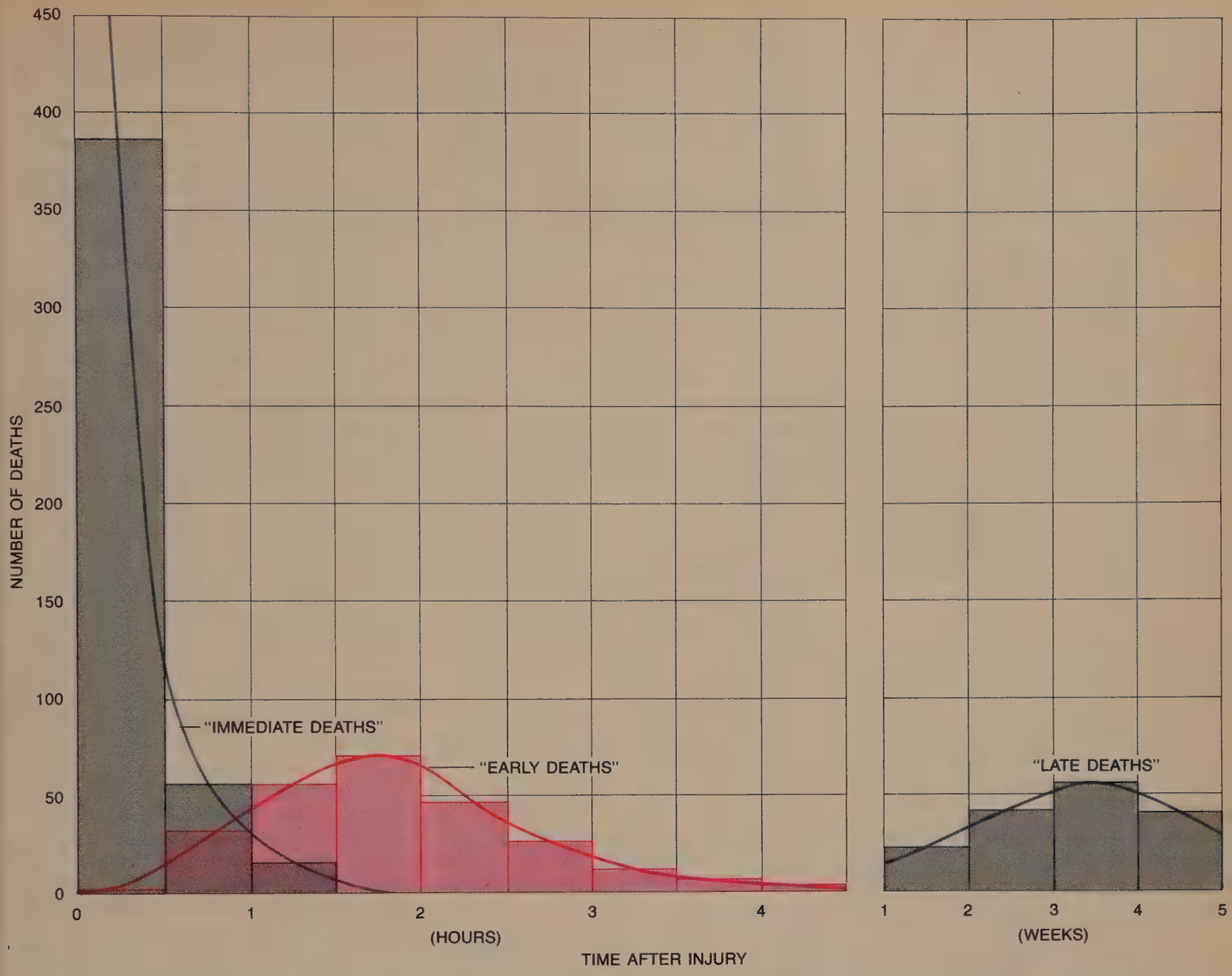
Another vexing social issue with a bearing on the current rate of trauma focuses on the mandatory use of safety devices such as automobile seat belts and motorcycle helmets. Legislation requiring the use of seat belts has been introduced in at least 20 countries. The results of these measures vary, depending on the degree of enforcement and compliance. So far the best record has been achieved in Australia, where after a law was passed requiring the use of seat belts there was a 27 percent decrease in the death rate from motor-vehicle accidents. Mandatory seat-belt legislation has not been popular in the U.S., however, in part because of the active resistance of groups opposed to such forms of Federal regulation.

The situation is equally unsettled in the case of laws requiring the use of motorcycle helmets. Beginning in 1967 a Federal highway-safety standard required that all states enact and enforce motorcycle-helmet laws. In the next decade fatalities from motorcycle accidents decreased by 50 percent nationally. Then in 1976 Congress revoked the sanctions against states not in compliance with the Federal standard. Over the next three years 27 states repealed or weakened their motorcycle-helmet laws. The result so far has been a 40 percent increase in the death rate from motorcycle accidents in those states. A recent study sponsored by the National Highway Traffic Safety Administration concluded that "the use of a safety helmet is the single most critical factor in the prevention or reduction of head injury" from motorcycle accidents.

The burden placed on society by unhelmeted motorcyclists can be demonstrated. In one study of 71 motorcyclists admitted to Denver General Hospital it was found that only 38 percent were covered by commercial medical insurance or workman's compensation; most of the unpaid bills were borne by the taxpayers. Similarly, in a survey done at the Maryland Institute for Emergency Medical Services it was found that 40 percent of the 65 motorcyclists hospitalized did not have insurance coverage.

Perhaps the most controversial trauma-prevention issue in the U.S. is that of handgun control. According to advocates of greater restrictions on the availability of handguns, more than 11,000 of the 26,000 murders recorded in the U.S. in 1982 were committed with such weapons. In addition, suicides and accidents involving handguns accounted for at least 10,000 deaths. Both figures are extraordinarily high, particularly in comparison with other industrialized societies, where handguns are controlled. There are at present some 60 million handguns in the U.S., and it can be argued that it would be difficult, if not impossible, to eliminate them entirely, even if the political will to do so were to prevail in Congress. In the meantime other approaches to handgun control, such as the institution of mandatory jail sentences for the criminal use of a handgun, might help to reduce the death rate from gunshot wounds. Laws of this kind have been passed in several states, and the results, particularly in Michigan, seem to be quite positive.

Another controversial issue is that of the decriminalization of narcotic drugs. In the 15 years that I have worked



TRIMODAL DISTRIBUTION of trauma deaths is observed when the death rate for a large enough sample of such deaths is plotted as a function of time after injury. The first peak (*"Immediate deaths"*) corresponds to people who die very soon after an injury; the deaths in this category are typically caused by lacerations of the brain, the brain stem, the upper spinal cord, the heart or one of the major blood vessels. The second peak (*"Early deaths"*) corresponds to people who die within the first few hours after an injury; most of these deaths are attributable to major internal hemorrhages or to multiple lesser injuries resulting in severe blood loss. The third peak (*"Late deaths"*) corresponds to people who die days or weeks after an injury; these deaths are usually due to infection or multiple organ failure. The graph is based on a sample of 862 trauma deaths recorded over a two-year period by the author's group at San Francisco General Hospital.

as a surgeon at the University of California at San Francisco General Hospital Medical Center the number of victims of penetrating trauma (primarily gunshot and stab wounds) has increased to approximately 40 percent of the total trauma caseload. Almost all of these injuries are related to drug trafficking. Most drug addicts must pay for their habit by illegal means, and violent crime is a common recourse. The decriminalization of drugs could help to solve at least this part of the drug problem.

Supporters of decriminalization argue further that the prevalence of drug abuse is not significantly dependent on the legal status of the drug in question. The experience of this country in the 1920's suggests that the consumption of alcohol was not reduced by prohibition; indeed, it may even have been increased. There is no reason to believe drug abuse would be greatly affected one way or the other by decriminalization. What is certain is that many of the negative social effects accompanying drug abuse would be alleviated.

Finally, there is the problem of burn injuries, which in many respects is representative of the larger trauma-prevention problem. More than two million Americans per year suffer from burns of one kind or another, and of them some 70,000 are admitted to a hospital. Of the latter group 8,000 or so eventually die of their burn injuries. More than a third of these deaths are attributable to cigarette smoking. The average American cigarette contains additives in both the paper and the tobacco that cause the cigarette to burn for approximately 28 minutes. If these additives were omitted, the cigarette would burn out in less than four minutes. As it happens, most furniture, upholstery and mattresses made in the U.S. need more than four minutes' exposure to a burning cigarette for ignition. The problem and the solution are obvious. Omitting the incendiary additives from cigarettes would not change the taste of the cigarette smoke, but it would make smoking safer by reducing fire-related deaths, disabilities and property losses.

Of course, the cigarette manufacturers are not about to remove these additives voluntarily. That change undoubtedly calls for Federal legislation. Just as in the case of motor-vehicle accidents caused by drunk driving, fires caused by cigarette smoking bring death and disability to innocent people as well as to the individuals responsible for the acci-

dents. Beyond the question of controlling cigarette additives, other burn-prevention proposals call for measures to promote such practices as the manufacture of flame-resistant clothing, the installation of smoke alarms, the shortening of cords on appliances and the reduction of water-heater temperatures to prevent scalds of small children.

Although the preceding discussion is not a complete catalogue of trauma-prevention issues, it does indicate some of the problems such proposals face in the U.S. Clearly these issues are complex and impinge on long-established social customs. Nevertheless, prevention remains the only feasible way to reduce the toll of immediate deaths from trauma. It is not only the most effective way to save lives but also the cheapest: crisis intervention after the fact is always expensive. Ultimately prevention could also help to reduce the other two death peaks from trauma, topics I shall now address.

Roughly 30 percent of the deaths from trauma fall into the category of early deaths. This category can in turn be subdivided into two major pathological conditions: neurological injuries and various kinds of hemorrhage. According to a recent nationwide survey of head and spinal-cord injuries, head injuries account for about .2 percent of all hospital admissions in the U.S. On this basis one can calculate that roughly 34,000 cases of traumatic intracranial bleeding are treated annually in the U.S.

The results of another recent study, done by a group at the Health Sciences Division of Virginia Commonwealth University, point to the need for prompt management of such head injuries. The Virginia group found that if surgical intervention for intracranial bleeding was delayed for more than four hours after an injury, the most probable outcome was death or permanent disability. If definitive surgical care was provided within four hours after an injury, however, the likelihood of a favorable outcome was significantly enhanced.

The need for prompt, definitive surgical intervention is also critical in the treatment of patients with injuries resulting in hemorrhage. For the sake of discussion hemorrhage can be divided into three grades: severe, moderate and minor. In cases of severe hemorrhage the rate of blood loss exceeds 150 milliliters per minute. In the first 10 minutes of severe hemorrhage the patient will lose at least 1,500 milliliters of blood, or roughly a third of his blood volume. If this rate continues unchecked, the patient will lose more than half of his blood volume within 20 minutes of the injury. In such cases little can be done in the prehospital setting to control the hemorrhage. Prompt, definitive surgical care offers this patient his only chance of survival.

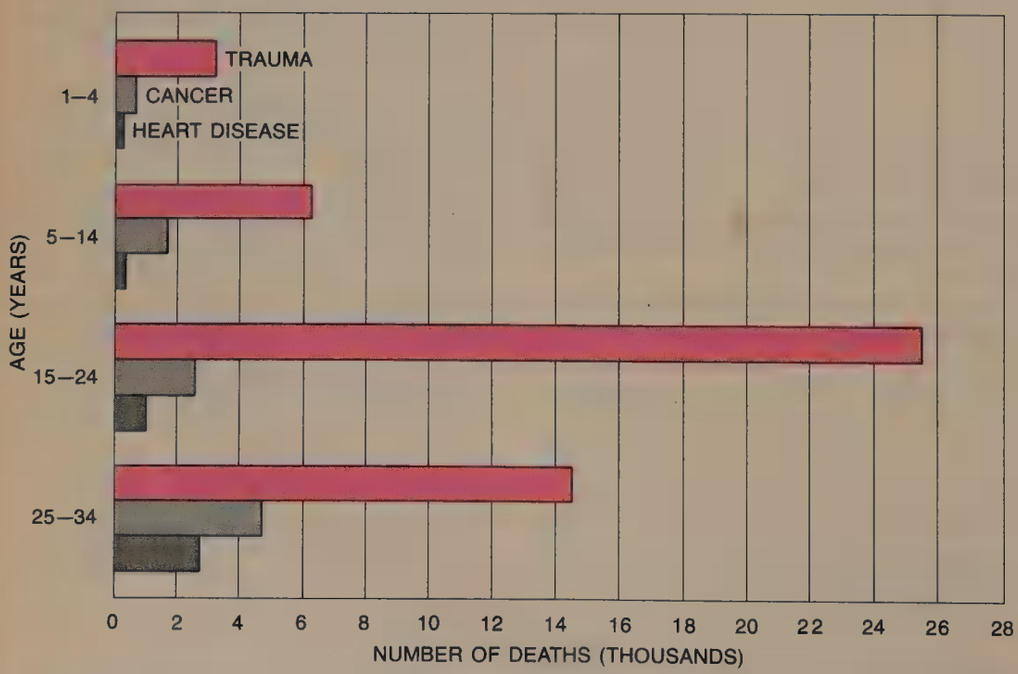
In cases of moderate hemorrhage the bleeding rate is between 30 and 150 milliliters per minute, and there will be a life-threatening blood loss within an hour of the injury. Rapid transport of the patient to a place where prompt surgical intervention is available is also the preferred treatment. Patients with minor hemorrhage (bleeding rates of less than 30 milliliters per minute) may have the "luxury" of an hour or more before surgical intervention is necessary. In addition intravenous lines started in the prehospital setting may keep up with the bleeding. In any case the main point remains that many early deaths from trauma could be prevented by rapid transport of the injured patient to a trauma center.

Are trauma patients in the U.S. receiving the kind of timely medical care these studies indicate they need? Several sources show that with the exception of a few communities having modern trauma centers the answer is no. For example, in 1960 investigators from the Surgeon General's office in Texas examined the deaths of 606 soldiers injured in accidents and treated in community hospitals. They found that 96 of the patients would have survived if appropriate treatment had been administered in time, and an additional 103 patients might have been saved if they had been treated appropriately. Another study, done by workers at the University of Michigan in 1969, showed that 28 of 159 patients who died as a result of trauma were inadequately treated. Still another study, reported by a group at Johns Hopkins University in 1972, showed that a third of the deaths resulting from motor-vehicle accidents involving abdominal injuries in the Baltimore area could have been prevented by prompt surgical intervention.

I have been personally involved in several studies of this kind in the San Francisco Bay area. The first study, reported in 1974, compared the death records from the trauma center at San Francisco General Hospital with those from several community hospitals in the surrounding area. The results showed that patients with injuries from motor-vehicle accidents treated in a hospital without a trauma center had a significantly greater chance of dying than those treated in the one with a trauma center. A subsequent study compared deaths caused by motor-vehicle accidents in one part of California where there was no trauma center (Orange County) with those in a part of the state that had a single designated trauma center (San Francisco). Again the outcome was significantly worse in the region without a trauma center.

The latter finding led to a follow-up study, initiated by physicians in the Orange County area. The data were reevaluated by an independent group of general surgeons, neurosurgeons and emergency-room physicians. Their report showed that non-neurological trauma care was inadequate in the hospitals without a designated trauma center. As a consequence five trauma centers were established in the Orange County area in 1980.

Another pertinent study was recently completed by the same group. It showed that the preventable-death category in Orange County dropped from 73 percent to 4 percent when patients were tak-



THREE LEADING CAUSES OF DEATH among young Americans are compared. The mortality figures, compiled by workers at the National Center for Health Statistics, are for 1977.

en to a trauma center rather than to a conventional hospital. Furthermore, the group found that none of the patients in the study died as a result of bypassing a conventional hospital in order to get to a trauma center. This finding emphasizes the importance of regional trauma care. Numerous other studies in various parts of the country lend further support to the conclusion that injured patients taken to hospitals without a trauma center are at a marked disadvantage. On the average the incidence of preventable deaths resulting from inadequate trauma care was found in these studies to vary between 30 and 40 percent.

Perhaps the most comprehensive study of this kind was done in 1980 by Daniel K. Lowe and his associates at the Oregon Health Sciences University. The study included 23 hospitals in a six-county region around Portland. The region has a total extent of 5,724 square miles and a population of 1,257,450, distributed over urban, suburban and rural areas. Originally 763 trauma patients were enrolled in the study; of these 104 had minor injuries and were excluded from further consideration. Of the 659 remaining patients there were 105 cases of inappropriate care, as determined by an independent trauma panel composed of general surgeons, neurosurgeons and emergency-room physicians. Of the 278 deaths registered 135 occurred in the hospital, and of these 34 were judged preventable by the panel. The latter group included 15 patients with brain injuries and 19 with various kinds of hemorrhage.

Of particular interest in the Portland results was the finding concerning the response time of the surgical consultants. In general the surgical consultants (there were 304 in all) took an average of 1.26 hours to get to the hospital after being called into the case. Neurological consultants responded somewhat more promptly: the average was .98 hour. The independent panel considered delayed response to be a significant factor in some of the cases of inadequate care. This finding draws attention to another problem: the popular misconception that any physician can treat a trauma patient adequately in a hospital emergency room. The emergency-room physician can start resuscitation, but a surgeon is almost always needed to provide definitive care. The sooner this care is provided, the better the outcome will be.

These findings all lead to one conclusion: There is a major shortfall in the delivery of trauma care in the U.S. The number of preventable deaths resulting from the existing system (or nonsystem) of trauma care is clearly unacceptable. What can be done to organize a better system of trauma care in the U.S.?

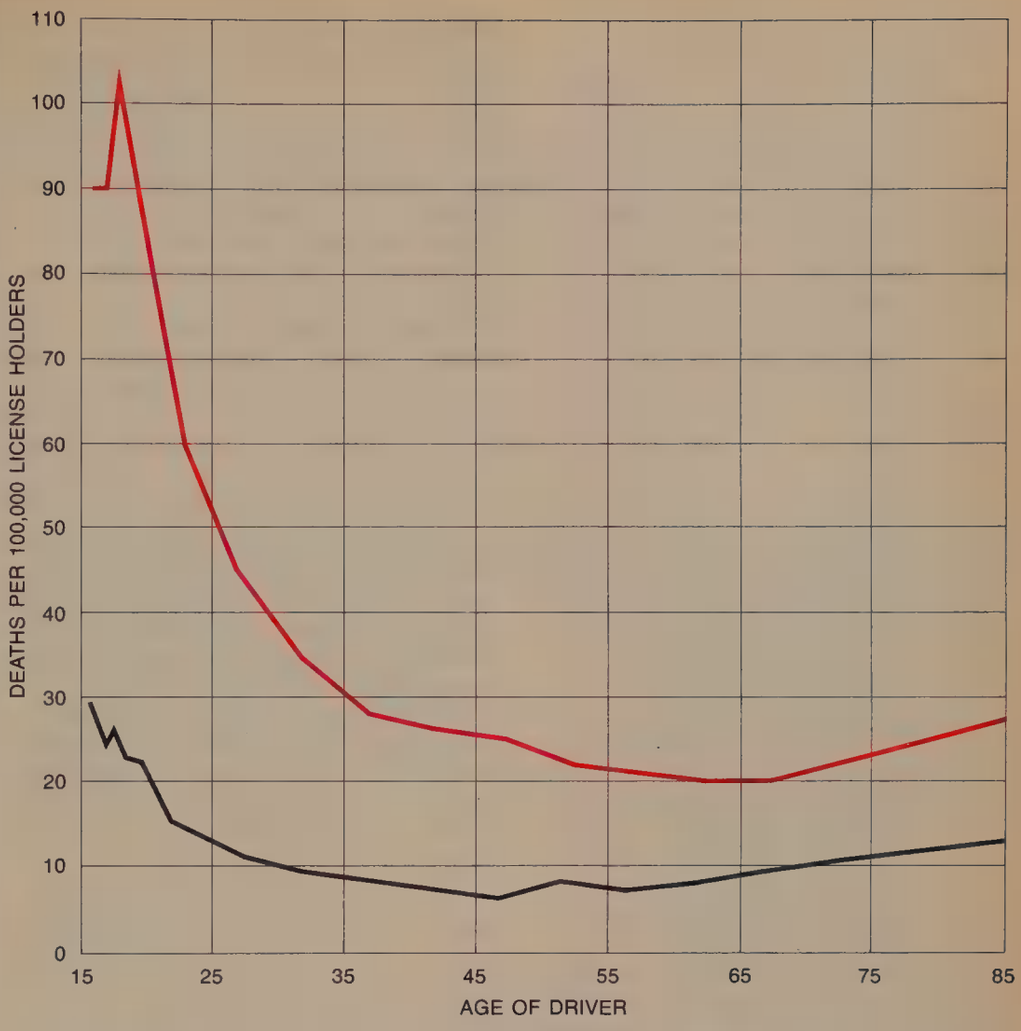
The concept of organized trauma care is not a new one; indeed, it can be traced back in military history to antiquity. The earliest mention of organized battlefield care is in the *Iliad*. According to Homer, Greek soldiers wounded in the fighting for Troy were carried off the battlefield and cared for in barracks (called *klisiai*) or on nearby ships. The *Iliad* contains references to 147 different wounds and implies a mortality rate of 77 percent.

The Romans also had considerable experience with emergency care for the injured; as early as 480 B.C. wounded soldiers were reportedly assigned to the care of patrician families. By the first century A.D. special hospitals (called *valetudinaria*) had been established along the borders of the Roman Empire to care for wounded legionaries. Archaeologists have identified at least 25 of these structures, which were quite sophisticated in design.

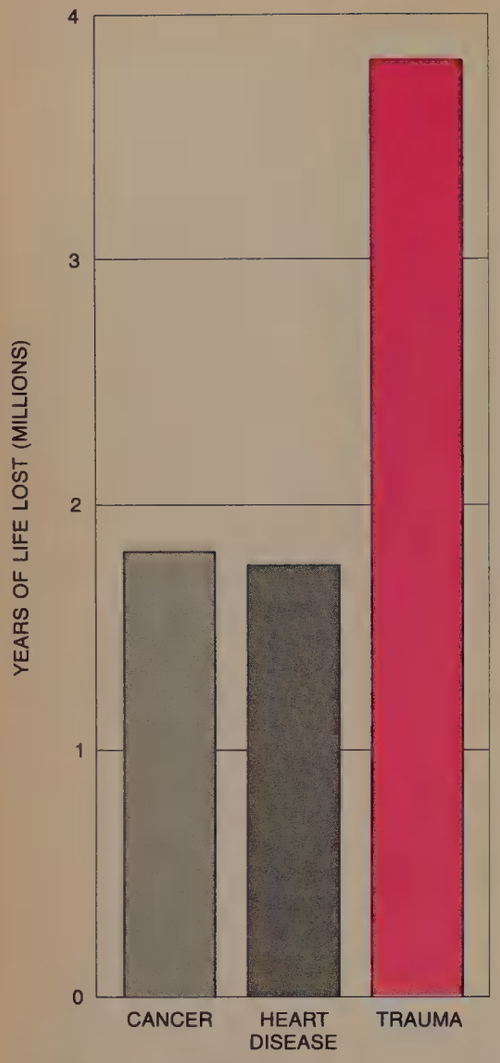
In the early 19th century Baron Dominique Jean Larrey, the chief surgeon in the army of Napoleon, made two improvements in the care of wounded soldiers that have persisted to modern times. The first was the *ambulance volante* ("flying hospital"), an innovation that sharply reduced the time it took to provide definitive care to the wounded. Before the advent of Larrey's horse-drawn ambulances injured soldiers often lay on the battlefield for periods of a day or more. Larrey's second innovation was to concentrate the casualties in one area and to operate on them as close to the front lines as possible.

During World War I the time lag between injury and surgery was still between 12 and 18 hours, and the overall mortality rate was 8.5 percent. The time lag was reduced in World War II to between six and 12 hours, and the mortality rate fell to 5.8 percent. Perhaps the most dramatic reduction in the time lag from injury to definitive surgical treatment came during the Korean conflict. A decision was made in the U.S. Army Medical Corps to bypass the battalion aid station and to take injured soldiers directly from the battlefield to a mobile army surgical hospital (M.A.S.H.). The average time lag between injury and definitive care during the Korean conflict was between two and four hours, and the mortality rate was 2.4 percent.

This tactic was further improved on during the U.S. involvement in Vietnam, where casualties were taken directly from the battlefield to the corps surgical hospital. According to one study, the average time lag between injury and de-



YOUNG MALE DRIVERS account for a disproportionately high percentage of deaths from motor-vehicle accidents, as is shown in this graph of the death rate for male drivers (*color*) and female drivers (*black*). Curves, for 1978, are from the Insurance Institute for Highway Safety.





finitive surgical care was reduced to 65 minutes, and the mortality rate fell to 1.7 percent. This military experience, one might think, should have served as an incentive and a model for the improvement of civilian trauma-care systems. With the exception of a few isolated instances, however, that has not happened.

One example of excellent regional trauma care can be found in West Germany. During the late 1960's West German health officials observed the U.S. methods of providing battlefield care in Vietnam. In 1970 the decision was made to apply these procedures throughout most of West Germany, establishing trauma centers along the main autobahns. Integral to the trauma-center concept in West Germany is rapid prehospital transport, which primarily entails the use of helicopters but also includes ground vehicles. There are now 32 air-rescue stations in the country with a standard mission radius of 50 kilometers. It is estimated that 90 percent of the population are within 15 minutes of a trauma center.

More important than the prehospital system is the system of integrated trauma care within the hospitals. West German hospitals have been classified according to their ability to provide trauma care. Furthermore, there is an in-house team of surgeons in every designated trauma center on a 24-hour basis. The teams include not only surgical residents but also a chief surgeon. The other important members of the trauma team are a neurosurgeon and an anesthesiologist. The chief trauma surgeon also cares for the patient in the postoperative period, including the time spent in the intensive-care unit. Overall the quality of surgical care is excellent.

The West German system also has a strong rehabilitation program, the primary goal of which is to get the accident victim back to gainful employment as soon as possible. I do not mean to imply that the system there is perfect. Some of the trauma centers are not as strong as they should be, but in general the system is an excellent model for the U.S.

As a consequence of this regionalized system the mortality rate from motor-vehicle accidents in West Germany has dropped from 16,000 per year in 1970 to 12,000 per year at present, a reduction of 25 percent. It is probably more than a coincidence that the magnitude of this reduction is remarkably close to the preventable-death estimate made in most American studies (between 30 and 40 percent). By applying simple arithmetic and assuming that the 4,000 additional German patients who now survive each year return to work, a rough estimate of the financial benefit to that society can be made. If one assumes that each survivor over the past 10 years now earns the equivalent of $10,000 per year and pays $2,500 in taxes, the gross national product of West Germany would be increased by $220 million per year and tax revenues would rise by $55 million. The value of a trauma center, therefore, lies not only in a reduction in deaths and disabilities but also in a positive financial contribution to society. If the U.S. were to introduce a similar system and could achieve the same reduction in mortality, then over the first 10 years this country's G.N.P. could be increased by more than $2 billion and the additional taxes paid would amount to more than $550 million.

The final category in this discussion, late deaths, accounts for approximately a fifth of all trauma deaths. Of these deaths 80 percent are attributable to infection and multiple organ failure. The two conditions seem to be causally related. The common risk factors that have been identified so far include shock, head injury, peritoneal contamination and malnutrition, all of which can lead to infections late in the course of a patient's injury. This development may in turn be related to the failure of the patient's immune system, but the exact causes have not been elucidated. Once infection is obvious the patient often develops progressive organ failure. The resulting mortality rate is high and is directly related to the number of organ systems involved.

The answer to the question of why the trauma patient is at risk for infection and multiple organ failure can only come from further research. Even this solution, however, is not without difficulties. At present the U.S. spends very little of its research funds on trauma. National priorities are clearly directed to cancer and heart disease, even though trauma accounts for more years of life lost than cancer and heart disease combined. One solution would be to establish a National Institute of Trauma, on the model of the National Cancer Insti-

MISMATCH between the cost of trauma, in terms of the number of years of life lost, and the national effort to solve the trauma problem, in terms of dollars spent on research, is particularly striking in the context of a comparison with the corresponding figures for cancer and heart disease. The bars in the top chart are based on an estimate published in the Surgeon General's report for 1975. The bars in the bottom chart are based on 1982 figures from the National Institutes of Health; they refer only to research funds spent under the auspices of the National Cancer Institute, the National Heart, Lung, and Blood Institute and (in the case of trauma research) the National Institute of General Medical Sciences.

tute or the National Heart, Lung, and Blood Institute. Such an institute could serve many purposes. It could focus on trauma both as a medical issue and as a social one. It could approach trauma in such a way as to place equal emphasis on prevention, health-care delivery and research aimed toward the solution of the late-deaths problem. It could also serve as a focus for innovative ideas in research funding. For example, since drunk driving is a major contributor to the national trauma rate, perhaps it would be feasible to devote part of the tax on alcoholic beverages to help support trauma research.

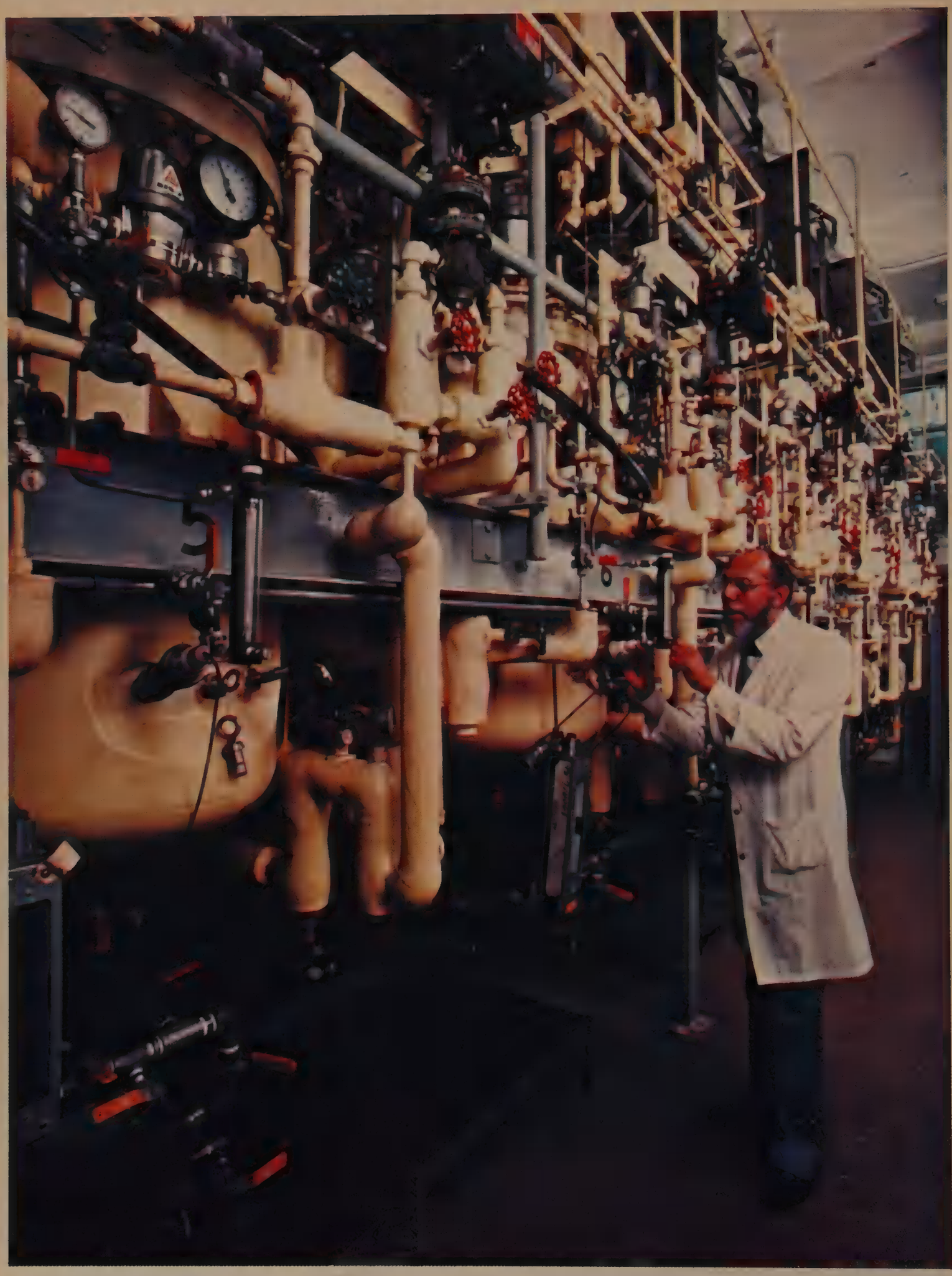
Finally, there is the problem of rehabilitation. One of the most pronounced deficiencies in trauma care in the U.S. is the lack of an integrated rehabilitation system. Most disabling injuries are caused by neurological and orthopedic injuries. With the exception of some excellent rehabilitation centers for spinal-cord injuries the U.S. has not placed enough emphasis on returning the injured patient to work. This approach should involve not only physical rehabilitation but also job retraining and treatment of the emotional trauma that often accompanies physical trauma. In order to have an effective rehabilitation system the nation must also address some of the existing worker's compensation laws and disability-reimbursement programs. For example, an employed person or union member currently takes five times as much disability time as a person who is self-employed. Any proposed rehabilitation system must address the disincentives that affect the decision to return to work.

In summary, trauma is a seriously neglected public-health problem in the U.S. Each of the three peaks in the trimodal distribution of trauma deaths has its own set of associated problems. If the U.S. is to achieve a significant reduction in the rate of mortality and disability from trauma, each of these issues must be addressed vigorously. The solutions will not be easy, and they will inevitably engage some controversial social issues. It is my contention, however, that the U.S. can no longer afford the present rate of preventable death and disability resulting from trauma. The search for solutions to the trauma problem must become a national priority.





MODEL SYSTEM for the delivery of trauma care has been established in West Germany. In the past decade or so specialized trauma centers have been set up at hospitals along the main autobahns. Most patients are transported to the designated trauma centers by helicopter. The red dots on the map at the right designate the system's principal air-rescue stations; the red circles mark the standard 50-kilometer operating radius associated with each station. According to West German health officials, 90 percent of the population are now within 15 minutes of a trauma center. The smaller version of the map reproduced at the left is on a sticker distributed by the German automobile club ADAC. Telephone numbers accompany names of the stations; an alternative emergency number (*110*) is also given.



INTERFERON IS MANUFACTURED by bacteria in this battery of 400-liter fermentation tanks at Hoffmann–La Roche, Inc., in Nutley, N.J. *Escherichia coli* cells into which recombinant DNA carrying a gene for human alpha interferon has been inserted are grown in a culture medium. The bacteria proliferate, synthesizing the protein interferon along with thousands of their own proteins. When the bacterial cells reach maximum concentration, they are killed, discharged from the fermentation tanks and concentrated by centrifugation.

The Purification and Manufacture of Human Interferons

Their promise is still not fulfilled, but now their genes have been isolated and cloned in E. coli. The bacteria are making interferon in quantity and the bacterial product is undergoing clinical trials

by Sidney Pestka

In the 1930's several investigators described the phenomenon of viral interference, whereby the infection of an animal by a virus seemed somehow to protect it against subsequent infection by another virus. In 1957 Alick Isaacs and Jean Lindenmann of the National Institute for Medical Research in London found an agent of viral interference: a protein, released by cells exposed to a virus, that enables other cells to resist viral infection. They called it interferon.

The great promise of interferon as an antiviral agent was evident from the moment of its discovery, which was supported by an independent report of similar findings by Y. Nagano and Y. Kojima of the Institute for Infectious Diseases in Tokyo. The promise was implicit in the fact that the protein is not directed against any one virus but rather protects cells against a wide range of viruses. It does more than that, affecting a number of different cellular activities in ways that suggest therapeutic possibilities. It is a potent substance: a very little of it goes a long way. As a natural cell product it seemed likely (given the correct dose level) to be safer than most new experimental drugs.

Yet the original promise has still to be fulfilled. For some 20 years after the discovery of interferon new complexities and difficulties seemed to turn up as often as promising leads. For one thing, interferon soon was found to be not a single protein but a large family of proteins, varying from species to species and present in multiple forms even within a species. Its mode of action proved to be indirect and is still not clearly understood. Above all, it is secreted by cells in minute amounts and was extremely difficult to purify. For many years it was hard to accumulate enough interferon for effective clinical trials. What was available was a crude preparation containing some interferons and a large amount of other protein. In the absence of purification it was impossible to pin down the specific activities of individual interferons, to determine their structure and so differentiate clearly among them, and to assess their safety and efficacy.

In the past few years the situation has changed. First a number of human interferons were purified, making it possible to begin to understand their structure and to categorize their activities. Then, very quickly, the availability of recombinant-DNA technology led to the isolation of human-interferon genes, their cloning in bacteria, the production in quantity of recombinant human interferon by fermentation and its purification by means of monoclonal antibodies. The first trials to test dose levels and the side effects of the purified bacterial product in human beings began in 1981. Now extensive tests of interferon's efficacy against viral diseases and cancers are under way. There has been a remarkably rapid transfer of new biological technology from the laboratory to pharmaceutical trials.

Isaacs and Lindenmann discovered interferon made by chick cells exposed to influenza virus and found it protected other chick cells, but not the cells of other animals, against infection by viruses. Whereas interferon is not virus-specific (in keeping with the concept of viral interference), it seemed to be species-specific, with each animal species manufacturing its own interferon. Then things got more complicated: it was found that a species makes several kinds of interferon, each of which has a particular spectrum of activity in other species.

Here I shall discuss only the human interferons, of which three classes have been well defined. For historical reasons they have long been designated as leukocyte interferon, fibroblast interferon and immune interferon. Now it is known that a particular class is not made by a single cell type, and so a new nomenclature has been proposed: alpha (for leukocyte), beta (for fibroblast) and gamma (for immune) interferon. Even this is an oversimplification. Many different members of the alpha class have already been isolated and characterized. There are reports that the beta class may have several members. So far only one member of the gamma class has been isolated and purified.

The reason interferon's protective effect is not limited to a single virus is that (unlike an antibody) it does not interact with the attacking virus; interferon interacts with the cell it protects. The manufacture of interferon is induced when a virus introduces its genetic material (DNA or RNA) into the cell. The presence of the foreign material (probably in the form of a double strand of RNA) causes the cell to synthesize and secrete molecules of interferon. The secreted interferon binds to a specific re-

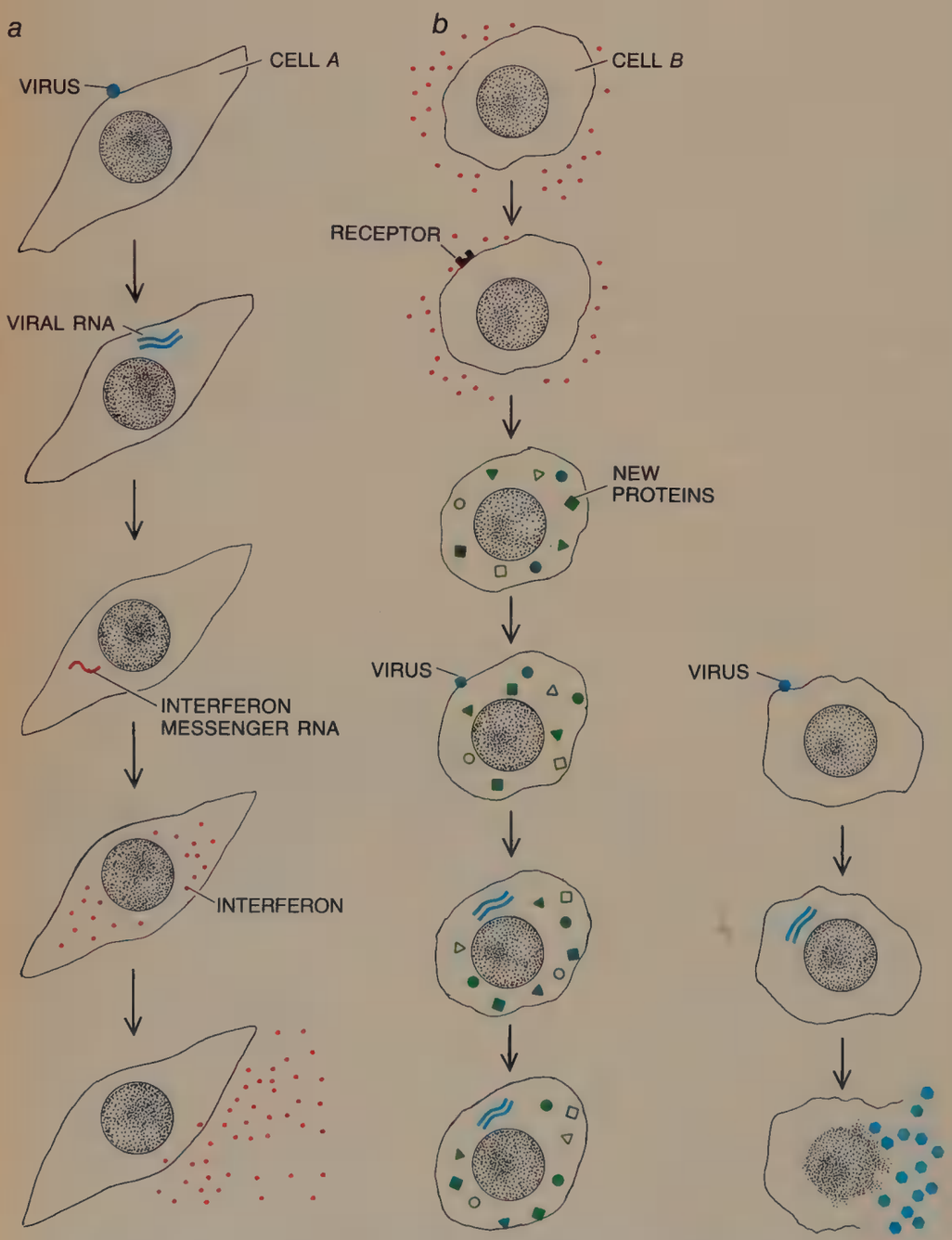


PASTE OF E. COLI is removed from the centrifugation vessel. The packed cells are then resuspended and broken open; cellular debris is separated by centrifugation. Nucleic acids and other viscous materials are removed and proteins in extract are passed through monoclonal-antibody column to purify interferon.

ceptor on the surface of other cells and in doing so apparently triggers several different mechanisms within these cells, initiating the synthesis of a number of proteins (some of which have been identified). The proteins somehow cause the cells to resist the usual effects of viral infection: multiplication of the virus, lysis (bursting) of the cell and liberation of progeny viruses. Much remains to be learned about interferon's mode of action, but the experiments of Thomas C. Merigan, Jr., of the Stanford University School of Medicine and others have established that even crude preparations have some clinical effect against certain viral infections.

As I mentioned above, interferon has multiple biological activities other than those contributing to the antiviral effect. Two are of particular interest. The protein appears in many cases to inhibit the proliferation of cells. It also somehow stimulates the activity of certain cells of the mammalian immune system: the lymphocytes called natural killer cells, which have a role in the destruction of foreign cells and perhaps of cancer cells. Either or both of these effects could explain evidence reported over the years, notably by Hans Strander of the Karolinska Institute, that interferon may promote the regression of some tumors.

Both the sporadic cancer studies and those attempting to assess interferon's efficacy against viral diseases suffered from the serious shortage and high cost of interferon and even more from the fact that the major "interferon" available was really a mixture of various proteins of which less than 1 percent by weight was interferon itself. Aside from antiviral activity no observed effect of the mixture could dependably be attributed to interferon. Purification of the protein was therefore a matter of high priority.

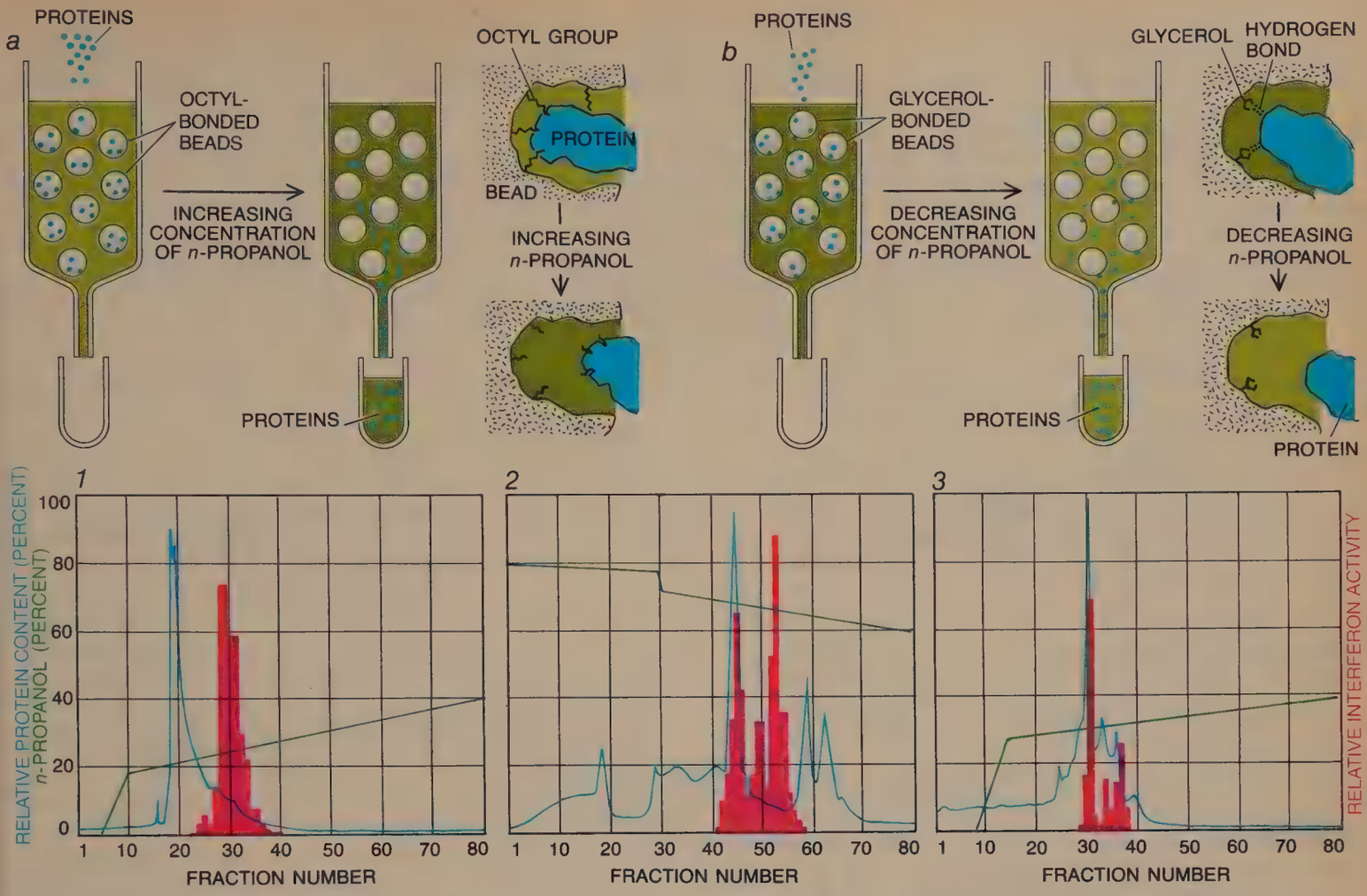


SYNTHESIS AND ANTIVIRAL ACTIVITY of interferon proceed in stages. The protein is made and secreted (*a*) by a cell infected by a virus. It is apparently the presence in the cell of double-strand viral RNA that gives rise to the synthesis of interferon messenger RNA, which is thereupon translated into interferon. The secreted interferon binds to specific receptor molecules on the surface of another cell (*b*). The binding of the interferon triggers a number of changes in cellular activity, including the synthesis of proteins that render the cell resistant to infection by a virus, which in an unprotected cell (*right*) would replicate and burst the cell.

In 1977 we took on the task of purifying human alpha interferon in my laboratory at the Roche Institute of Molecular Biology. The first requirement was a large supply of crude leukocyte interferon. Our method of preparing it was essentially the one developed several years before by Kari Cantell of the Finnish Red Cross Blood Center. Human leukocytes (white blood cells) are incubated with an inducing virus, either Sendai virus or Newcastle-disease virus. Where Cantell had used human or bovine blood serum as a component of the culture medium, we substituted the milk protein casein; as a single protein it proved to be easier to remove in the initial concentration steps than the multiple proteins of serum were. We found the yield could be improved by substituting leukocytes from patients with chronic myelogenous leukemia for normal leukocytes, and we were able to obtain large supplies of leukemic cells (which are removed in therapy of the disease) from the M. D. Anderson Hospital and Tumor Institute in Houston.

After incubation overnight the cells and viruses are removed by centrifugation. What remains is a crude interferon preparation: it contains some induced interferon along with any other proteins whose secretion may have been induced by the virus and also all the normal cell secretions. The objective is to remove all the proteins other than interferon in a series of purification steps. At each step one must assay the concentrate for interferon activity in terms of standard units of "specific activity." The usual assay, measuring the extent to which a sample inhibits the destruction of cells by a virus, took three days. Philip C. Familletti and Sara Rubinstein found a way to cut the time to less than 16 hours and thereby speeded up the purification process considerably.

Knowing how much difficulty had attended efforts to purify interferon by conventional techniques, we decided to try what is called high-performance liquid chromatography. Chromatography methods in general involve adsorbing a crude mixture to some solid support and eluting different fractions with a solvent. In high-performance liquid chromatography the starting mixture is adsorbed to very fine beadlike particles packed in a column and the solvent is pumped through the column. Sidney Udenfriend, Stanley Stein and Peter Böhlen of Roche had managed to



INTERFERON WAS PURIFIED by means of high-performance liquid chromatography (*top*). In the reverse-phase method (*a*) silica beads to which octyl groups have been bonded are packed into a chromatography column; proteins pumped into the column bind tightly to the octyl groups. Increasing concentrations of the solvent *n*-propanol in an aqueous solution are passed through the column; the *n*-propanol releases different proteins from the column in the order of their increasing affinity for octyl groups. In the normal-phase method (*b*) glycerol is bonded to the beads. In the presence of a high *n*-propanol concentration, proteins pumped into the column form hydrogen bonds with the glycerol. As the *n*-propanol is diluted proteins are released in the order of their increasing ability to form hydrogen bonds with the glycerol. Purification data (*bottom*) show how reverse-phase chromatography (*1, 3*) was alternated with normal-phase chromatography (*2*). The protein content of successive fractions was monitored and each fraction was assayed for interferon activity (*red bars*). Interferon-rich fractions were pooled for further purification. Eventually a pure interferon fraction was isolated (*3*).

separate peptides (short chains of amino acids) by means of reverse-phase liquid chromatography, in which the solid phase (the beads) is coated with an organic material that is hydrophobic, or water-repellent; the mobile phase (the solvent) is more polar, or water-attracting. Stein, Menachem Rubinstein and I undertook to apply the method to purify the alpha interferon.

Ethyl alcohol, the accepted solvent, did not work; the protein remained adsorbed to the beads. We decided to try a somewhat less polar solvent, *n*-propanol, even though proteins are not very soluble in it and the interferon might precipitate as it was being eluted. As it turned out, a gradient of *n*-propanol worked. As we pumped increasing concentrations of the solvent through the column, different protein fractions (having different affinities for the solid support) were eluted successively and collected in tubes. Each fraction was assayed for interferon activity. Fractions that were relatively rich in interferon were applied to another column for further purification. By alternating the reverse-phase process with normal-phase chromatography (in which the beads are coated with hydrophilic groups and the solvent is less polar) Rubinstein purified human alpha interferon, in just a few steps, about 80,000-fold. The specific activity of the purified interferon was from 200 to 400 million units per milligram. Subjected to gel electrophoresis, which separated the proteins on the basis of molecular size, the interferon yielded a single band at a molecular weight of 17,500, and the protein in that single band was active. In other words, human alpha interferon had been purified to homogeneity.

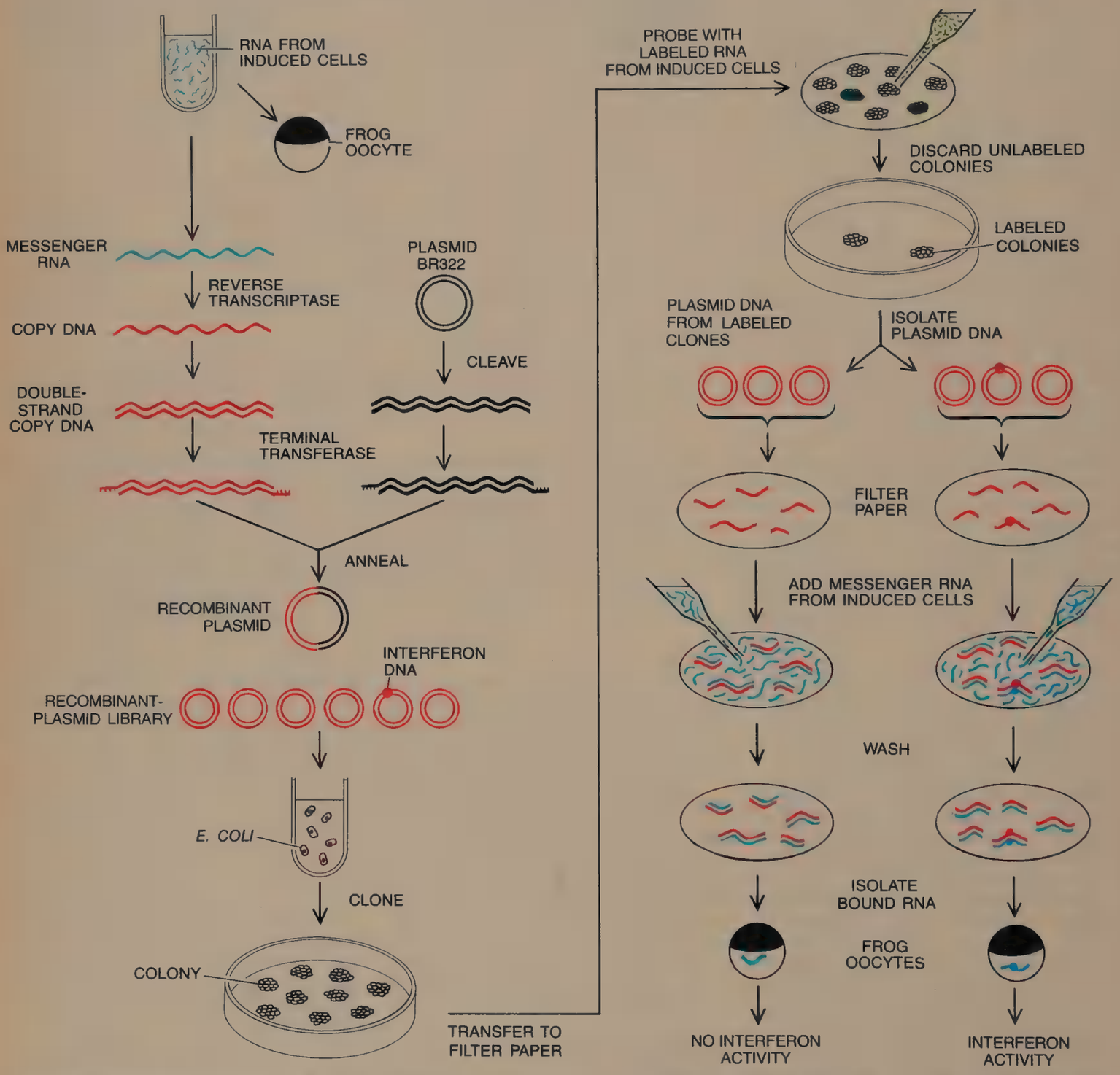
During the purification process we had been surprised to note that interferon activity was purified in several different fractions, indicating that we were isolating a number of different species of the protein. Chemical characterization of the purified products showed this was indeed the case. When we treated the proteins with trypsin, an enzyme that cleaves the long chain of amino acid subunits at specific sites, different purified products yielded somewhat different sets of peptides. When we had enough of each of several purified interferons, we proceeded to analyze their amino acid composition (the number of each of the 20 different amino acids present) with a sensitive amino acid analyzer Stein had built. Again there was evidence for the presence of multiple species.

There had been earlier reports that different crude preparations of human alpha interferon were somewhat different in the electric charge of the molecules, but this had not been attributed to variations in the molecules' amino acid sequence, the ultimate determinant of a protein's individuality. Interferons were thought to be glycoproteins (proteins to which sugar chains are attached), and the charge heterogeneity was thought to be due to differences in carbohydrate content. John A. Moschera looked at five purified alpha species and found no detectable carbohydrate in any of them, however. Soon Christian B. Anfinsen's group at the National Institute of Arthritis, Metabolism, and Digestive Dis-

eases and Geoffrey Allen and Karl H. Fantes of the Wellcome Research Laboratories in Britain isolated multiple alpha interferon species; Allen and Fantes looked for carbohydrate and did not find any. Some interferon species may indeed turn out to be glycosylated, but the dogma that all interferons are glycoproteins is invalid.

With the amino acid compositions in hand, several groups went on to tackle the actual amino acid sequences. Kathryn C. Zoon and her colleagues in Anfinsen's laboratory, collaborating with Michael W. Hunkapiller and Leroy E. Hood of the California Institute of Technology, got the first partial sequence, for the amino terminus (the beginning) of their human alpha interferon. A few months later Warren P. Levy of my laboratory and John E. Shively of the City of Hope Medical Center determined the amino-terminal sequence of one of our alpha species. At two sites Zoon's sequence and ours showed a different amino acid. The particular amino acids involved were not likely to be mistaken for each other, and so we felt both sequences must be correct. Here was more clear evidence that the alpha interferons are a family of closely related proteins, a finding that was confirmed when Levy and Shively, Zoon and her colleagues and Allen and Fantes eventually obtained more sequences.

Although our alpha interferon was

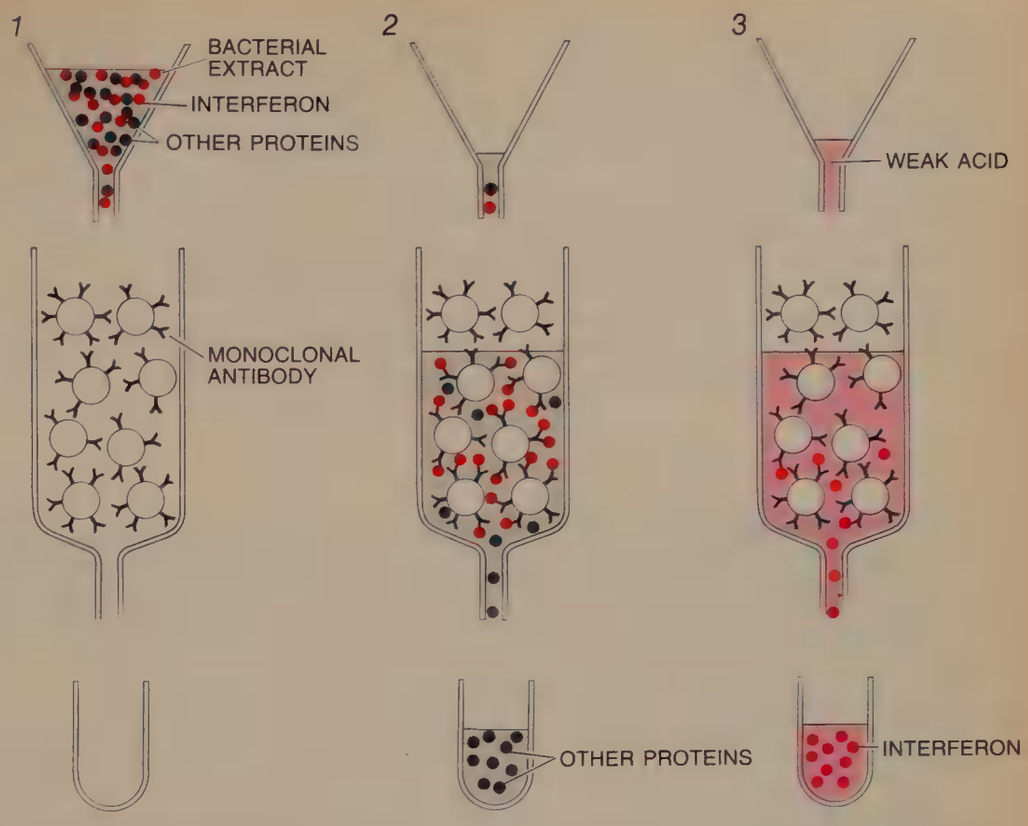


CLONING OF INTERFERON GENE begins with the induction of interferon synthesis in leukocytes. Messenger RNA from the cells is assayed for interferon activity in frog oocytes. If a sample shows interferon activity, the RNA is copied into DNA; a second strand is made with DNA polymerase. The "copy DNA" (cDNA) is treated with terminal transferase to produce "sticky ends." Meanwhile a vector, plasmid BR322, is cut open and complementary sticky ends are synthesized. The cDNA is annealed with the plasmid DNA to make a "library" of recombinant plasmids, which are introduced into *E. coli* cells. Clones of identical cells are grown from individual *E. coli*.

To find a clone harboring plasmids carrying interferon DNA the colonies, replicated on filter paper, are probed with radioactively labeled RNA from induced cells under conditions such that the probe hybridizes with, and thus identifies, DNA encoding induced proteins, including interferon. The plasmid DNA of the labeled colonies is cut into linear fragments, batches of which are fixed to filter paper. RNA from induced cells is added; it hybridizes with DNA encoding induced proteins. The hybridized RNA is separated from the DNA and assayed in oocytes to identify plasmid-DNA batches encoding interferon. Procedure is repeated for individual clones of positive batches.

the first interferon to be purified in such a way that its composition could be determined, Ernest Knight, Jr., of E. I. du Pont de Nemours & Company had purified some human beta (fibroblast) interferon a few years earlier. Now Knight, Yin H. Tan of the University of Calgary and Stein, Heinz-Jürgen Friesen and Moschera in my laboratory worked out the terminal sequences of purified beta interferon. The sequences were all the same; so far only a single beta interferon has been isolated by each group, and the proteins all seem to be identical.

With homogeneous human alpha and beta interferon available it was finally possible to assess the biological activity of pure interferons rather than crude mixtures. The antiviral activity of interferons turns out not to be strictly species-specific. Most of the human alpha (leukocyte) interferons are about as active in protecting bovine cells as they are in protecting human cells; two of them are much more active on bovine cells than on human cells. Human beta (fibroblast) interferon has very low antiviral activity on bovine cells, however. Working with our purified alpha species, Marian Evinger in my laboratory and Nathan O. Kaplan of the University of California at San Diego found that antiproliferative activity too is a property of interferon itself, but that some versions of the protein are much better than others in this respect. As for the stimulation of killer-cell activity, Ronald B. Herberman and John R. Ortaldo of the National Cancer Institute found that almost all the pure interferons do enhance the ability of certain lymphocytes to destroy target cells.



MONOCLONAL ANTIBODIES are exploited to purify recombinant interferon. A crude interferon mixture extracted from bacteria is poured onto a column packed with beads to which monoclonal antibodies against alpha interferon have been linked (*1*). The interferon binds to the antibodies and remains attached as all other proteins pass through the column (*2*). A weak acid dislodges the interferon (*3*); the solution is neutralized and the interferon is concentrated.

There remained the problem of obtaining enough pure interferon for intensive study of its mode of action and for clinical trials. We might have accumulated enough by large-scale induction of leukocytes and purification of the crude material, but a better method was at hand. In the mid-1970's recombinant-DNA techniques had been developed whereby the DNA containing the gene for a particular protein can be inserted into the bacterium *Escherichia coli* and be cloned, or replicated in many copies, and expressed: translated into protein. We realized that by such methods we could obtain a large supply of particular interferon genes for study and also get the bacteria to manufacture a particular interferon—eventually perhaps on an industrial scale.

Actually we had taken the first step a few years before. Ordinarily one begins the gene-cloning process by isolating, from cells that make a lot of the desired protein, the messenger RNA encoding the protein. (Messenger RNA is the nucleic acid into which the DNA of a gene is transcribed by the cell; it is subsequently translated by the cellular machinery to make a protein.) Interferon messenger RNA is present in leukocytes at a very low level, so that it is hard to identify. In 1974 James L. McInnes of my laboratory, working with Edward A. Havell and Jan Vilček of the New York University School of Medicine, had extracted messenger RNA from cells induced to manufacture interferon. They added the RNA to cell-free extracts containing the cellular translation machinery and succeeded in translating the RNA into active interferon. Later Ralph L. Cavalieri was able to insert the messenger RNA into frog oocytes (precursors of egg cells); the oocytes translated the messenger. By assaying the product of the cell-free system or of the oocytes for interferon activity one could determine whether or not a particular RNA preparation included at least some interferon messenger. M. N. Thang of the Institut de Biologie Physico-chimique and Paula M. Pitha of the Johns Hopkins University School of Medicine developed similar methods for measuring interferon messenger RNA.

Once Familletti had optimized the production of interferon messenger RNA in induced leukocytes, Shuichiro Maeda, Russell McCandliss and Alan Sloma prepared complementary DNA from the RNA and spliced the DNA into plasmids: small circles of bacterial DNA. This yielded a large "library" of recombinant plasmids, each carrying a DNA copy (cDNA) of some messenger RNA from the induced leukocytes. The plasmids were introduced into *E. coli* cells and individual clones (colonies descended from an individual cell) were isolated. The next problem—the hard part of the procedure—was to find, in the vast library of recombinant plasmids, those with DNA encoding interferon.

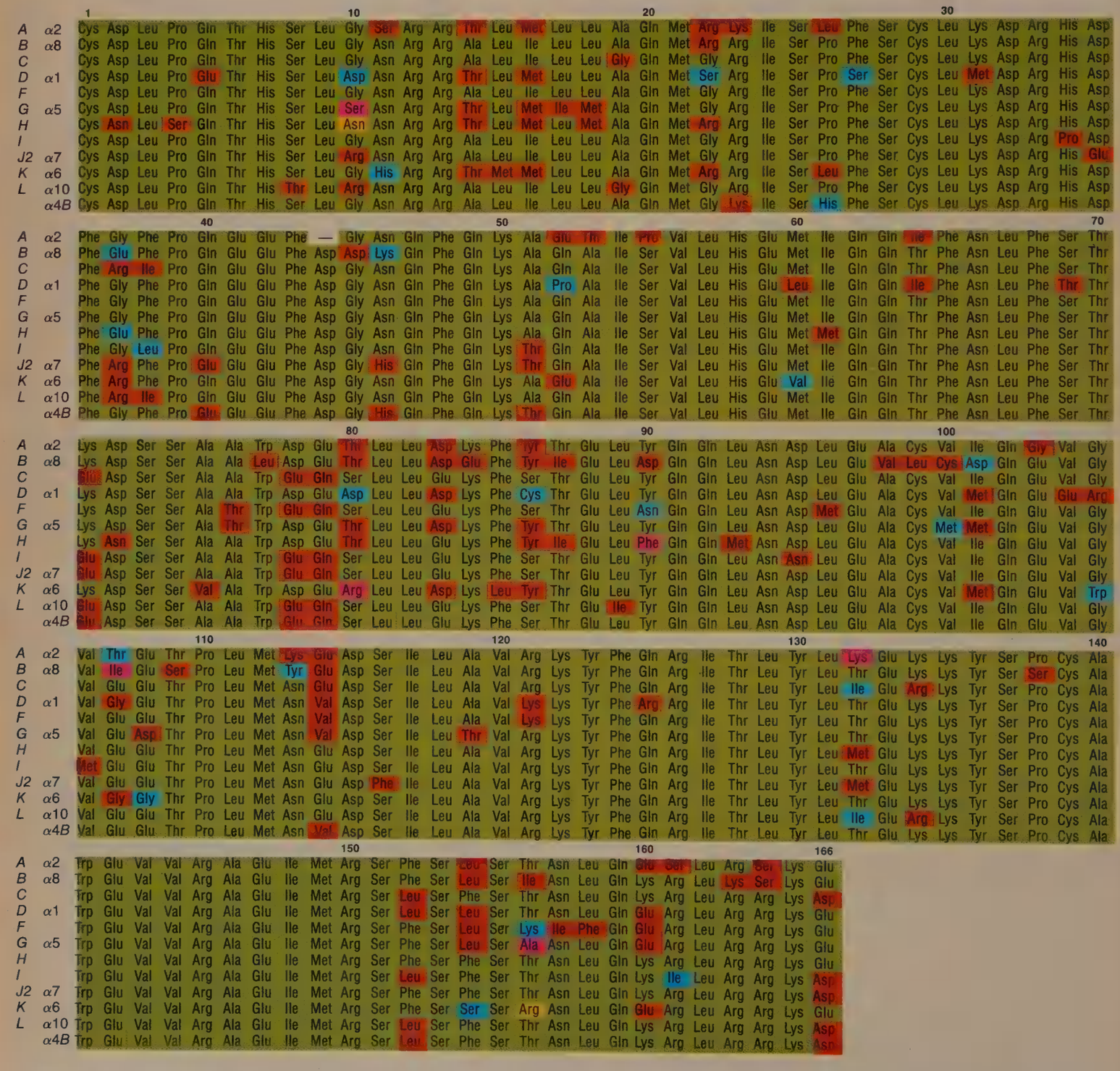
If we had been able to begin with a purified interferon messenger RNA, that would have been a straightforward task. DNA and RNA are chains of the subunits called nucleotides; the sequence of nucleotides encodes genetic information. RNA transcribed from a stretch of DNA is complementary to it in sequence and will hybridize with it, the two strands pairing to form a double strand. A pure interferon RNA can therefore be labeled with a radioactive isotope and used as a probe that anneals to the sought-after DNA, whose identity is revealed by autoradiography.

The reader will remember, however, that we did not have pure interferon RNA; we had begun with a mixture of RNA's whose interferon-specific activity could be revealed only after the fact, by translating the RNA and assaying the products for interferon activity. In this situation we had to devise an indirect, two-stage procedure. In the first stage we screened all the bacterial colonies to find those with plasmids carrying cDNA made from the RNA of induced cells and therefore perhaps carrying interfer-

on cDNA. We did it by probing the colonies' cDNA with labeled induced-cell RNA in the presence of large amounts of unlabeled RNA from uninduced cells, so that the labeled probe hybridized only with induced cDNA. The screening allowed us to discard some 90 percent of the colonies: those whose plasmids carried no induced cDNA and therefore could not encode interferon.

In the second stage we had to identify, in the remaining 10 percent of the colonies, particular recombinant plasmids carrying interferon cDNA. We did it by applying the messenger RNA from induced leukocytes to pooled batches of the plasmid DNA of the colonies. Any RNA that hybridized with the plasmid DNA was separated from the DNA and assayed in oocytes for interferon activity. Positive batches were then examined clone by clone in the same way.

By these means we found a recombinant, plasmid 104, that had most of the DNA sequence coding for a human alpha interferon. We found another, plasmid 101, carrying the sequence for human beta interferon. (This was quite a surprise, considering that its messenger RNA had been synthesized in leukocytes rather than in fibroblasts. By now it has become clear that a particular cell type may manufacture more than one class of interferon.) At about the same time several other investigative groups were reporting the cloning of human interferon DNA's. Shigekazu Nagata and Charles Weissmann of the University of Zurich and their colleagues had just



ALPHA INTERFERONS whose genes have been isolated form a family of closely related proteins. Here their amino acid sequences, deduced from the sequence of nucleotides in the DNA encoding them, are arrayed and colored to show the high degree of homology. Species whose genes were isolated by the author's laboratory in collaboration with Genentech, Inc., are indicated by the capital letters at the far left. Identical or closely related genes sequenced by Charles Weissmann's group at the University of Zurich are indicated by the *alpha* designations in the next column. The sequence displayed here is that of the author's version except in the case of α5 and α4*B*.

cloned an alpha-interferon DNA that was different from ours. The beta gene was cloned by Tadatsugu Taniguchi of the Japanese Foundation for Cancer Research (who had in fact reported success a few months before the rest of us), by Walter Fiers of the State University of Ghent, by Norman H. Carey and others at G. D. Searle & Co., Ltd., and also by us in collaboration with workers at Genentech, Inc.

Plasmid 104 served as a probe with which to search through a standard library of human DNA and find related interferon genes. So far we have isolated 16 distinct sequences for human alpha interferons; some of the same sequences and a few different ones have been isolated by other workers. None of the interferon DNA's contains intervening sequences, the noncoding stretches of DNA that interrupt protein-coding regions in most mammalian genes. The nucleotide sequences of the DNA's confirm what was indicated by the amino acid sequences of purified interferons: the alpha interferons are a class of related proteins, each one encoded by one of a family of related genes.

Plasmid 104 did not include the entire coding sequence for an alpha interferon; because of the vagaries inherent in the construction of recombinants the beginning of the gene was missing. At Genentech David V. Goeddel and his co-workers exploited plasmid 104 as a probe for searching through a library of cDNA until they found an entire gene. That gene needed to be reconstructed to achieve its expression in *E. coli.* First Goeddel removed from it a segment coding for the leader, a peptide that signals the cell to secrete the protein and is cleaved during secretion. Then Roberto Crea put together a segment of DNA carrying an initiation codon, a three-nucleotide signal to translate messenger RNA into protein; the segment was spliced to the gene just before the beginning of the segment coding for interferon's amino acids. Finally the Genentech workers attached, before the initiation codon, a regulatory sequence containing the bacterium's own signal to begin the transcription of DNA into messenger RNA. The result was a chimeric gene combining a bacterial regulatory region with the human gene's coding region for interferon.

When the reconstructed chimeric gene was inserted into *E. coli,* the bacteria proceeded to synthesize large quantities of human interferon: about as much per liter of bacterial culture medium as could have been produced by the leukocytes from 100 blood donors. Recombinant human leukocyte interferon *A,* as we designated it, behaves like its counterpart made from human cells. In an early animal experiment done by Noel Stebbing at Genentech it prevented a viral disease, encephalomyocarditis, in squirrel monkeys. Before the bacterial product could be tested in human beings, however, rigorous purification was necessary to remove contaminating bacterial proteins. We turned to monoclonal antibodies.

An antibody is a protein of the immune system that recognizes and binds to a foreign protein, or antigen. Since 1975 it has been possible to prepare large amounts of monoclonal antibodies: antibodies directed against a specific antigen. The availability of our purified interferon had enabled Theophil Staehelin of F. Hoffmann–La Roche & Co., Ltd., to prepare monoclonal antibodies directed against specific interferon molecules. Monoclonal antibodies that were specific for alpha interferon were linked to beads packed into chromatography columns. To purify bacterial interferon Staehelin, Hsiang-fu Kung and Donna S. Hobbs poured onto the column an extract of *E. coli* cells that had synthesized recombinant interferon *A.* The interferon, and only the interferon, bound to the antibodies; the other components, including any bacterial toxins, went right through the column. Then an acid solution was passed through the column to elute virtually pure interferon.

The availability of large amounts of very pure interferon opened the way to clinical trials, which I shall describe below. It also meant we could crystallize interferon, the first step toward analysis of the protein's three-dimensional structure by X-ray crystallography. David L. Miller and Kung have prepared crystals of human recombinant interferon *A.*

Early clinical trials of interferon were hampered, as I have suggested, by the short supply and high cost of interferon synthesized by human cells and by the impurity of the preparations, which clouded assessment of interferon's own effect and also limited dosages (since large amounts of the crude material could not be administered to patients). The purification of recombinant human alpha interferon in quantity removed all these impediments. After appropriate tests for safety in animals, the bacterial product prepared by Hoffmann–La Roche, Inc., was approved for trials in human beings. In January of 1981 Jordan U. Gutterman of the M. D. Anderson Hospital initiated a clinical trial designed to establish the safety, toxicity and side effects of various blood levels of recombinant interferon *A* in cancer patients.

More than 500 patients have been given interferon in trials conducted in various academic institutions and coordinated by Zofia Dziewanowska of Hoffmann–La Roche and her colleagues. The most frequent side effects were those noted earlier with crude interferon: fever, chills, muscle aches, mild gastrointestinal upset, fatigue and anorexia. Patients seemed to develop tolerance to acute "flulike" side effects, but fatigue and anorexia increased with dosage and duration of treatment. There was also some decrease in the white-blood-cell count, which reversed within a few days, and in a few instances there was some elevation of liver enzymes that seemed to have no deleterious effect.

Some of the studies were designed not only to assess tolerance but also to evaluate antitumor response. Most of the patients had advanced cancer resistant to conventional therapy. Even in this group some tumor regressions were observed in kidney cancer, malignant melanoma, multiple myeloma and a few other malignancies. Some results suggested that recombinant interferon *A* may be effective against Kaposi's sarcoma in patients with acquired immune deficiency syndrome (AIDS). Cancer of the breast, lung and colon appeared to respond only minimally or not at all. These are very preliminary results, to be sure. Additional trials of interferon *A*'s efficacy in cancer and viral diseases are under way.

The detection of cancers may also be improved by interferon. It was known to alter the protein composition of the cell surface. Recently Paul B. Fisher, I. Bernard Weinstein and Soldano Ferrone of the College of Physicians and Surgeons of Columbia University, Jeffrey Schlom and John W. Greiner of the National Cancer Institute and I have found that recombinant interferon *A* can increase the expression of certain tumor-associated proteins on the surface of malignant-melanoma, colon-cancer and breast-cancer cells. Such an effect may make it possible, with monoclonal antibodies to these tumor antigens, to diagnose a developing cancer many months earlier than is now possible.

Interferon *A* is only the first bacterial interferon to be tested. Other alpha interferons are becoming available for trials; beta interferon and gamma interferon will follow. Some classes, species and combinations of interferon may be more effective than others against particular diseases and under particular conditions. Moreover, the physician will not be limited to the set of natural interferons. One can break up interferon genes and splice the pieces to make new genes that are translated into hybrid interferons. We and other workers have already experimented with such hybrid molecules. It may become possible, as more is learned about the mechanisms of these proteins' various activities, to tailor interferon molecules to optimize particular effects.

Magnetic Fields in the Cosmos

The dynamo mechanism for the generation of magnetic fields explains why Venus has no field, why the sun has an oscillating field and why the dominant galactic field is parallel to the plane of the galactic disk

by E. N. Parker

If nuclear and gravitational forces were the only forces at work in the universe, the broad pattern of cosmic evolution would be one of gradual thermal degradation punctuated by occasional explosive events. The cosmos would resemble the serene—and monotonous—heavens of classical conception. There is, however, a cosmic agitator: the magnetic field. Although only a small part of the available energy in the universe is invested in magnetic fields, they are responsible for most of the continual violent activity of the cosmos, from auroral displays in the earth's atmosphere to stellar flares and X-ray emission, and the massing of clouds of interstellar gas in galaxies.

Within the solar system spacecraft have carried magnetometers near every planet from Mercury to Saturn. Mercury, the earth, Jupiter and Saturn have fields of their own, but Venus and Mars do not. The fields range in strength from 3.5×10^{-3} gauss at the poles (Mercury) to eight gauss at the poles (Jupiter). The earth's field, in comparison, is about .6 gauss at the poles. The *Voyager II* spacecraft, which carries a magnetometer, should pass close to Uranus in 1986. Magnetic fields are expected on both Uranus and Neptune because of their similarity to Jupiter and Saturn.

In all but a few cases the magnetic fields of stars cannot be detected directly from the earth, but their existence can be inferred from the presence of activity similar to solar magnetic activity. On such evidence it appears that most stars have magnetic fields at least as strong as the sun's. Some classes of stars do, however, have fields strong enough to be measured directly, for example by measuring the extent to which the two states of polarization of a given spectroscopic line are shifted from the position of the unpolarized line. Magnetic *A* stars have fields of up to 34,000 gauss; some white-dwarf stars have fields of from 10^7 to 10^8 gauss, and pulsars have fields of 10^{12} gauss. The disk of our galaxy is permeated by a field of between 2×10^{-6} and 3×10^{-6} gauss, and other galaxies appear to have fields at least as strong. There is even evidence, although it is controversial, that a weak magnetic field fills the space between galaxies.

Surprisingly, there is a single, generic explanation for the ability of bodies as different as a dense, cold planet and a tenuous, hot galactic disk to generate a magnetic field. The explanation, first worked out for the earth, comes from the discipline of magnetohydrodynamics: the study of the interaction of a moving, electrically conducting fluid and a magnetic field.

The cosmos is filled with fluids capable of carrying electric currents. Most of the fluids are hot ionized gases, but in planets they are internal reservoirs of molten metal. The energy released in the interior of planets and stars and in assemblages of stars by the action of nuclear and gravitational forces keeps these fluids in turbulent motion. The magnetic fields entrained in the fluids are stretched and folded by the fluid motion, gaining energy in the process. In other words, the turbulent fluids function as dynamos, devices that convert mechanical energy into the energy of magnetic fields. The ability of magnetic fields to feed (as living organisms do) on the general energy flow from the interior of planets, stars and galaxies explains why the fields flourish in and around nearly every celestial body.

Mathematical models of dynamos with various physical characteristics convincingly simulate the fields produced by astronomical bodies. The shell model of the sun's convective zone, for example, predicts a waxing and waning field that reverses polarity at regular intervals, as the sun's field does. Given the lack of detailed information about factors such as the motions of the gas within the sun, however, the dynamo models are inevitably less than definitive. In some instances the available information raises questions that cannot yet be answered. One example is the ability of Mercury to generate a field. Moreover, the dynamo mechanism by itself cannot account for the exceptionally strong fields of some stars. Because of such gaps in information, the rival hypothesis that there are primordial fields, trapped for example in the stable core of stars since their formation, cannot be disproved. The balance of the evidence, however, indicates that the planets, the sun, most stars and the galaxy function as colossal dynamos.

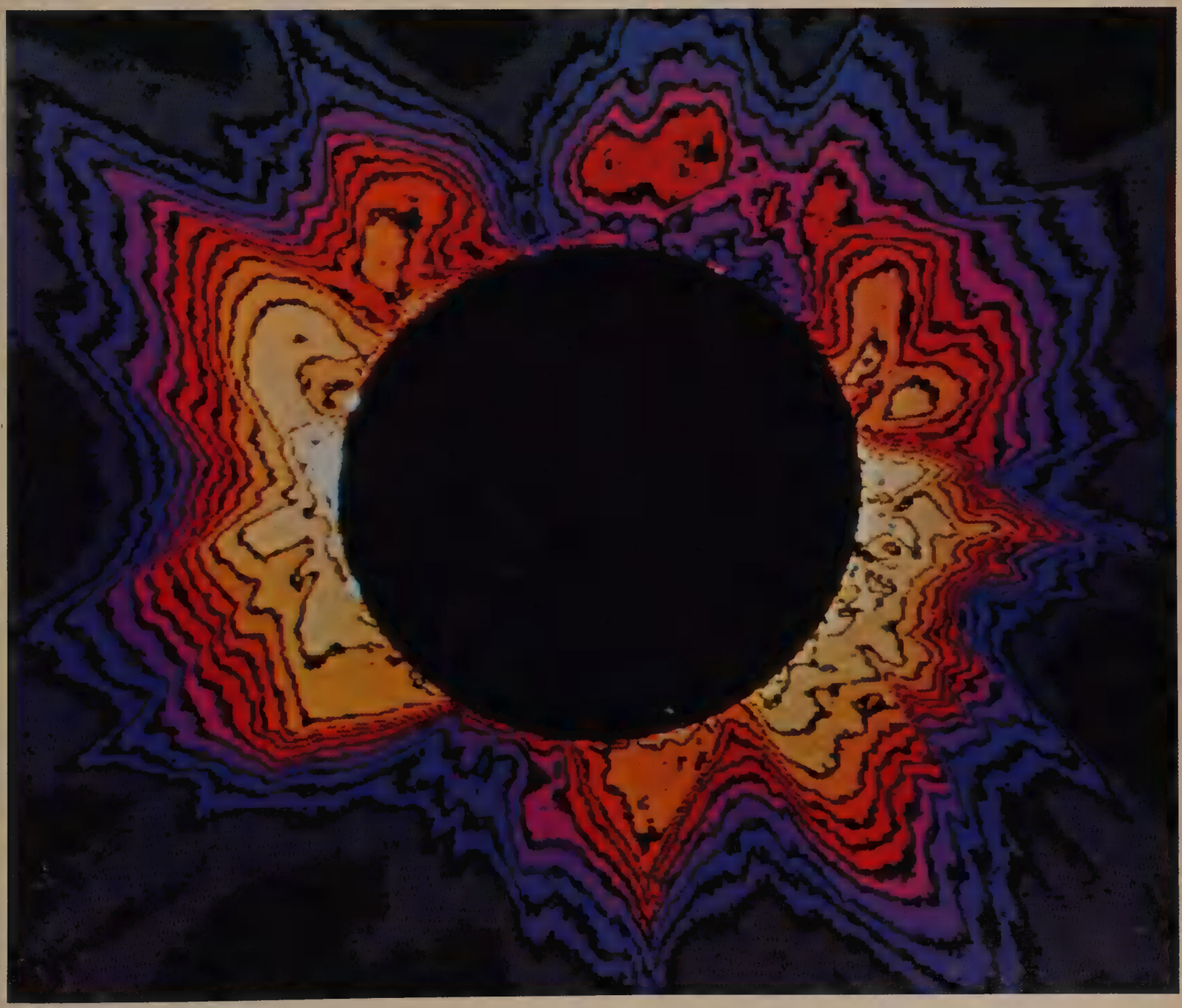
A dynamo converts the energy of motion of a conductor into the energy of an electric current and a magnetic field. A simple laboratory dynamo consists of a metal disk rotating on an axle and over a conducting coil the axis of which is aligned with the axle of the disk. The coil is electrically connected to the disk and the axle by brushes. A current passed through the coil induces a magnetic field aligned with the axis of the coil. The electrons in the disk moving through this field are subjected to a force at right angles to their direction of motion and the direction of the magnetic field, that is, a force directed along the radius of the disk. The direction of the force is given by the familiar right-hand rule. If the disk, viewed from above, is rotating counterclockwise and the magnetic field is directed upward, the induced electromotive force causes a current to flow from the axle to the periphery of the disk. The current then flows from the disk through the brush to the coil, amplifying the magnetic field generated by the coil, which in turn amplifies the current flowing through the disk. Magnetohydrodynamic dynamos operate on the same principle, although the electrons are not confined to coiled wires and can move throughout a body of fluid. Explaining how a moving conductive fluid amplifies a magnetic field becomes complicated, however, if the argument is pursued in terms of induced electromotive forces, the resulting currents and the magnetic fields associated with the currents.

Here some conceptual shortcuts are useful. The magnetic field can be visualized as consisting of lines of force, the closed loops along which a compass

needle would align itself. The strength of the field in any given volume of space can then be represented schematically by the number of lines that traverse the volume or, less conventionally, by the thickness of the lines. Following a point made by Hannes Alfvén, one of the founders of magnetohydrodynamics, one may regard the field lines as being "frozen" into the conducting fluid or "attached" to the particles of which the fluid is composed. The field moves with the fluid, and the field lines are deformed by the fluid motions as if they were rubber bands. If the particles to which a field line is "attached" are moving at right angles to the field line but at different speeds, the field line will be stretched. In accordance with such conventions the stretching corresponds to a gain in the strength of the field. The energy of the motion of the particles is converted into the energy of a magnetic field, and induced electromotive forces drive the current associated with the field just as they do in the disk dynamo.

It should be noted that the operation of either type of dynamo calls for the initial presence of at least a weak magnetic field or a weak current (the current in the coil in the case of the disk dynamo). Therefore the dynamo mechanism does not explain how the magnetic fields of planets and stars may have originated but rather how they are amplified and maintained in spite of the continual sapping of the field by the dissipation of the associated current. Other mechanisms that will be discussed below explain how the weak fields the dynamo mechanism amplifies could be created.

The ubiquity of magnetic fields suggests that the circumstances under which a magnetohydrodynamic dynamo develops are not particularly stringent. The first requirement is the presence of an electrically conducting fluid capable of supporting the currents associated with the field. Most of the universe is filled with such fluids. Most planets have a molten metallic core; stars are composed of ionized gases, and nearly all space is filled with a gas that has enough free electrons to be a good conductor of electricity. In most instances the gas is not as conductive as a



MAGNETISM OF THE SUN is apparent in a computer-graphics image of the sun's corona. Since the gas of the corona is ionized, the distribution of its electrically charged atoms reflects the form of the sun's magnetic field. This false-color image is a computer analysis of a photograph of the corona made in India during the total solar eclipse of February 16, 1980. The photograph was made by a research group from the High Altitude Observatory of the University of Colorado at Boulder and Southwestern at Memphis College with a special camera developed by Gordon A. Newkirk, Jr. In the camera a radially graded filter is interposed between the lens and the photographic plate. The filter blocks more light toward the center of the image than toward the periphery, preventing the bright inner corona from washing out the detail of the rest. The colors correspond to the intensity of the light at the wavelength of 6,400 angstrom units from each part of the corona; white areas have the highest intensity and deep violet areas the lowest intensity. Dark lines are contours of equal intensity.

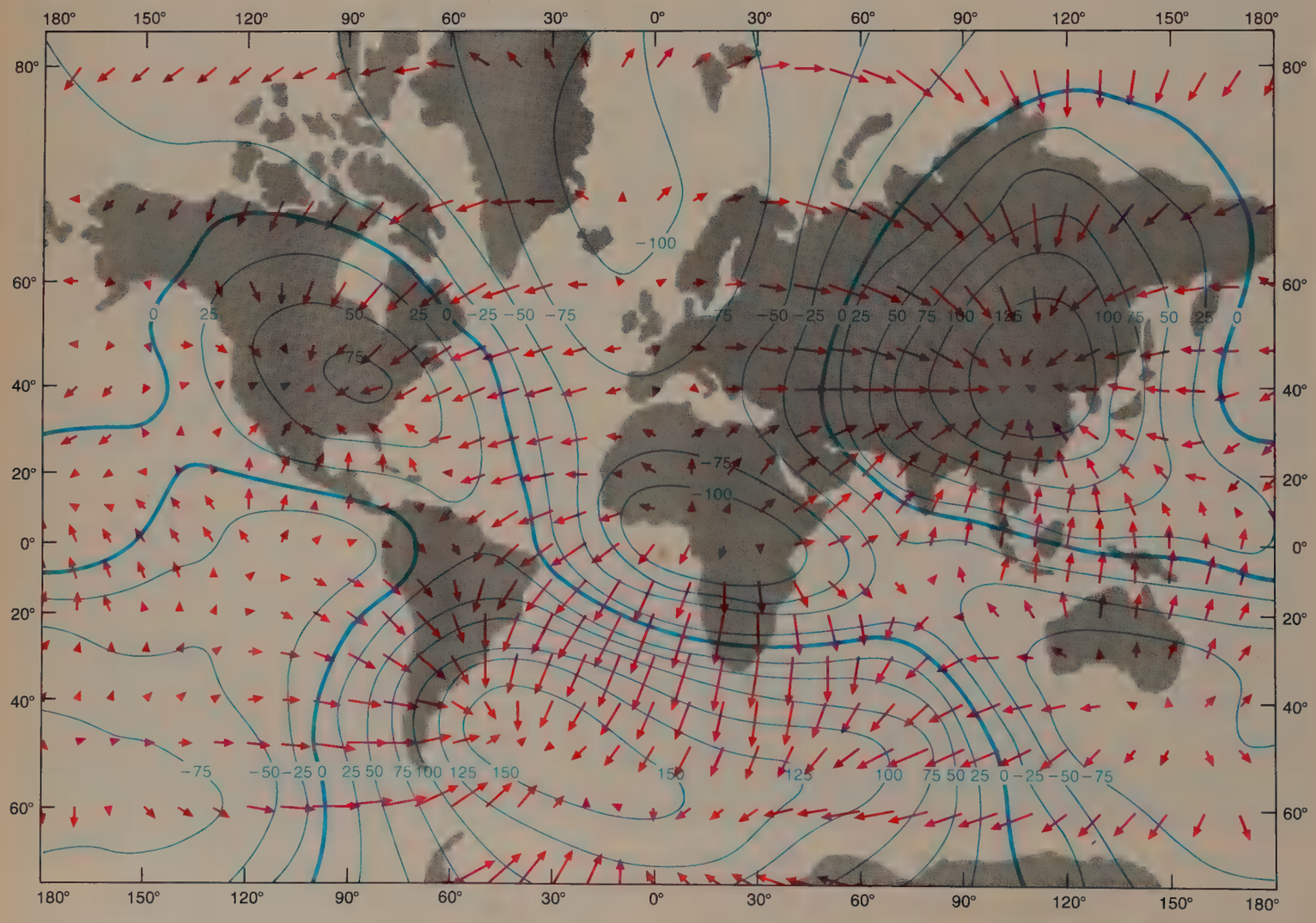
solid metal, but the current-carrying capacity of a body is proportional to its area as well as to its conductivity; hence the broad extent of the gas makes it capable of carrying large currents. The cold, dense, nonconducting atmosphere and the rocky, poorly conducting surface of planets such as the earth are the only effective electrical insulators to be found anywhere in the universe.

The second requirement is a pattern of fluid motion that amplifies the magnetic field. In principle many types of fluid motion would amplify the field, but the naturally occurring combination of nonuniform rotation and cyclonic convection is a particularly effective recipe. All that is necessary for such a combination is a rotating body either containing or composed of a convecting fluid. As a consequence of the convection the rotation of the fluid tends to be nonuniform, and as a consequence of the rotation the convection is cyclonic.

The most familiar example of this pattern of motion is the circulation of the earth's atmosphere. The variation of the direction of the prevailing winds with latitude reflects the nonuniform rotation of the atmosphere. For example, the tropical trade winds are easterlies because air whose angular momentum is determined at higher latitudes loses rotational speed as it moves farther from the earth's axis of rotation; hence in the equatorial zone the air is moving slower from west to east than the earth's surface. The atmosphere's convective cells (the high- and low-pressure regions) are given a cyclonic twist by the rotation of the earth. As a result rising (low pressure) areas spiral counterclockwise in the Northern Hemisphere and clockwise in the Southern Hemisphere; the reverse is true of sinking (high pressure) areas. The earth's atmosphere does not function as a dynamo simply because there are no free electrons to carry the necessary electric current. Similar motions can be expected, however, in the molten core of the earth, in the convective zone of the sun and on a gigantic scale in the gaseous disk of the galaxy, all of which are sufficiently conductive to function as dynamos.

The first cosmic dynamo to be successfully modeled was the terrestrial one. A necessary precondition for the development of the model was an accurate picture of the interior of the earth. This was supplied by the study of the paths and velocities of seismic waves spreading out from the focus of an earthquake. By 1940 it was known that the inner half of the earth's radius (the radius is 6,400 kilometers) is occupied by molten metal with an electrical conductivity not much less than that of ordinary solid iron. At the center of the liquid core there is a small solid core of crystalline metal with a radius about an eighth that of the earth. The small solid core plays no essential role in generating the magnetic field, and so it is ignored in most discussions of the subject. The outer half of the radius of the earth consists of the mantle and the crust. The hot silicate minerals of these layers are relatively poor conductors of electricity and for the present discussion can be treated as insulators. In this context the mantle is largely an impediment that hides the fluid motions and electric currents in the core and masks rapid small-scale fluctuations of the magnetic field itself.

The dipolar (two-pole) magnetic field detected at the surface of the earth is necessarily associated with circular electric currents flowing from east to



LOCAL ANOMALIES in the earth's magnetic field probably correspond to convection cells in the earth's core. These maps show the residual or irregular component of the earth's field, the part that remains when a north-south dipole field is subtracted from the total observed field. The arrows represent the horizontal component of the residual field. The longest arrows correspond to a field strength of about 10,000 gammas. (One gamma equals 10^{-5} gauss.) The contours connect points on the surface of the earth where the vertical

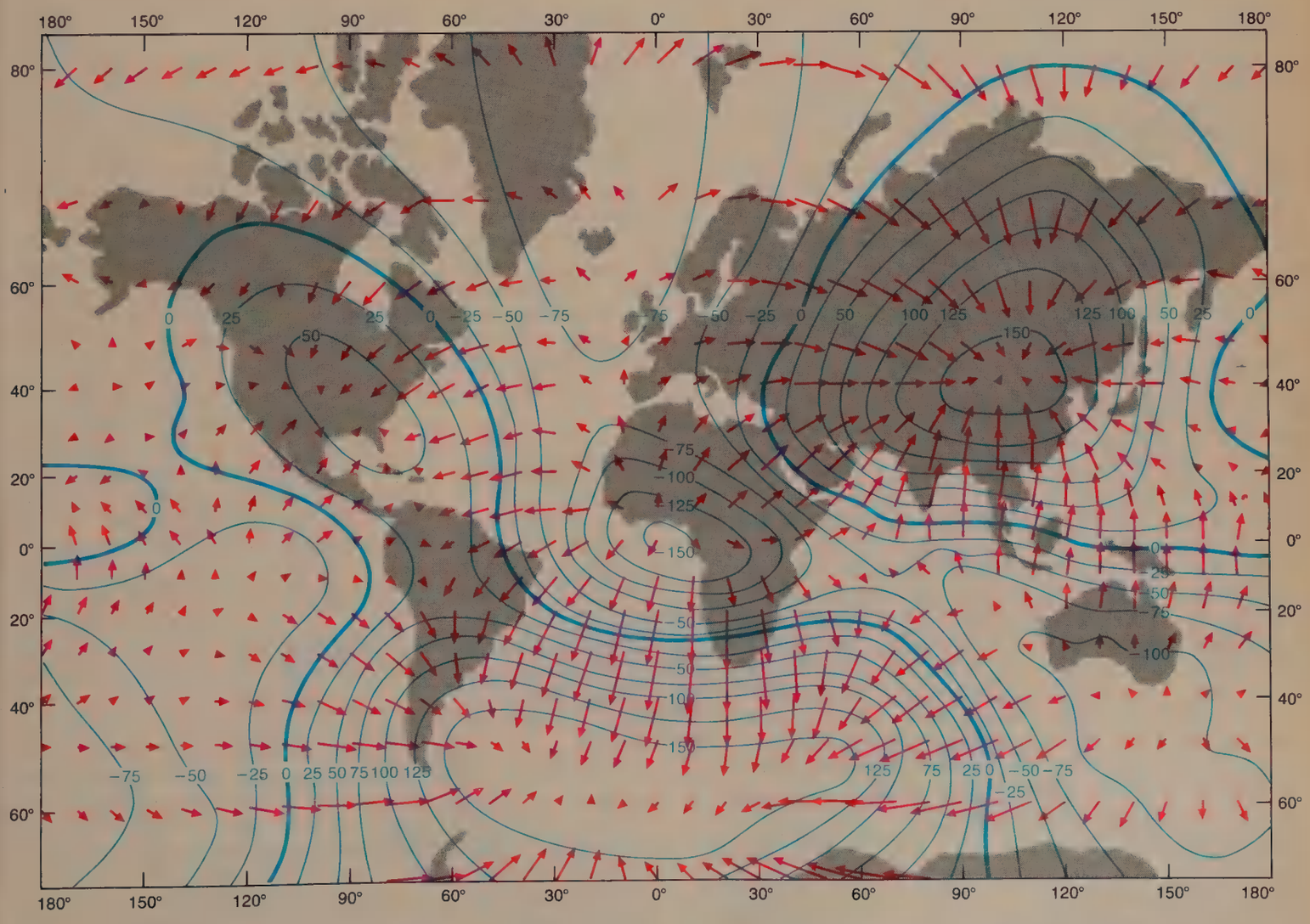
west in the molten metal core. A simple calculation indicates that associated with the field are currents of some 2×10^9 amperes. At first attempts to explain the generation of the earth's field appealed to a variety of atomic effects, such as the thermoelectric effect, as being the agencies potentially capable of producing the separation of charge required to cause the necessary flow of current. In 1945, however, Walter M. Elsasser pointed out that advances in atomic and solid-state physics made it possible to rule out the thermoelectric effect and all other atomic processes as the cause of the earth's field. He emphasized that the only remaining possibility was for the currents supporting the field to be induced by the motion of the conductive liquid across the lines of magnetic force of the field itself. He was proposing the dynamo effect.

The implication of this explanation of the earth's field is that the fluid outer core is in motion, something for which there is no direct evidence because of the interposition of the mantle. The variation of the earth's field with time, however, can be explained only by motions in the core. Records of the direction and magnitude of the field at the surface have been kept since the time of Carl Friedrich Gauss, the great 19th-century mathematician and physicist for whom the unit of magnetic-flux density is named. They show a dozen or more identifiable local anomalies in the field, anomalies that have dimensions of several thousand kilometers and amplitudes on the order of 10 percent of the dipole field. The anomalies change slowly with time, increasing or decreasing over lifetimes of several centuries. Elsasser pointed out that the changing magnetic pattern corresponds to a broad convective pattern of a dozen identifiable cells in the liquid core.

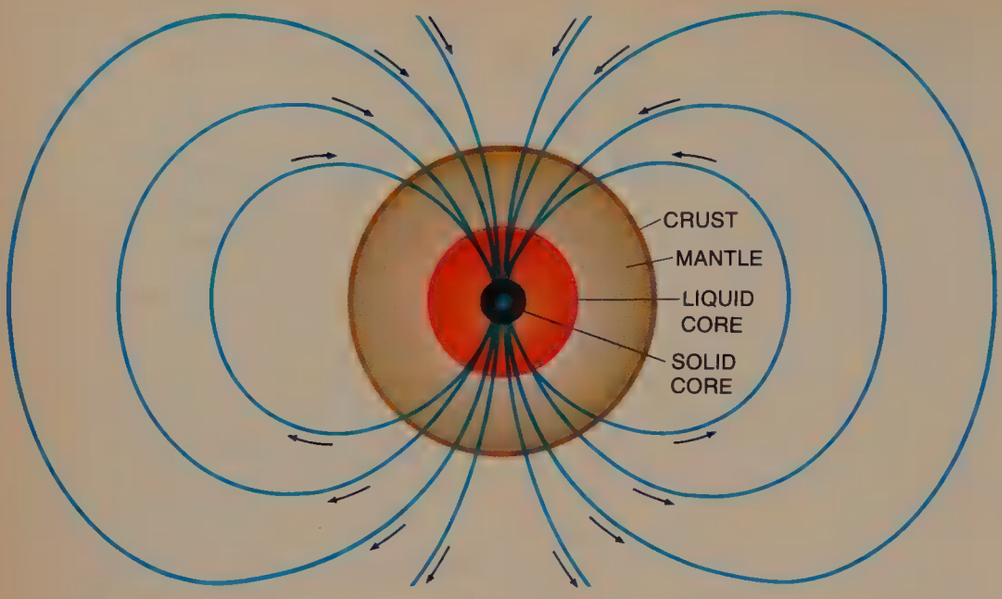
The known cells lie at low and middle latitudes. They have a definite tendency to drift westward at a rate of about .18 degree per year. The drift suggests that the surface of the core, through which the field emerges, rotates slower than the overlying mantle. The drift rate corresponds to a fluid velocity of .3 millimeter per second (about a meter per hour) at the surface of the core.

Elsasser noted that the most obvious explanation for the slow rotation of the core is the action of the Coriolis force on the rising and sinking fluid in the convective cells. The conservation of angular momentum requires that the angular velocity of the rising fluid decrease as it moves farther from the spin axis of the earth. It follows that the core surface rotates faster at high latitudes than it does at low latitudes and that at low latitudes the inner part of the core rotates faster than the surface.

Elsasser set out to calculate the effect of the assumed motion of the electrically conducting fluid in the core on the dipole magnetic field. The first important fact to emerge from his investigation was that the primary magnetic field in the core of the earth is an east-west field, at right angles to the main component of the field at the surface. It is called the azimuthal field; in general that part of a field lying in the planes perpendicular to the axis of rotation of a magnetic body is called an azimuthal field and that part lying in the planes through the axis is called a meridional field. The azimuthal field of the earth's core is created by the stretching of the north-south lines of force of the dipole field as they are carried around in the rotating fluid of the core. The part of a field line lying



component of the residual field is of equal magnitude. The contours are labeled in hundreds of gammas; negative numbers correspond to field directed downward from the horizontal. The magnetic anomalies drift westward at a rate of about .18 degree per year. For example, the magnetic anomaly in central Africa in 1907 (*map at left*) had drifted to the west coast of Africa by 1945 (*map at right*). The westward drift of the anomalies shows that the surface of the fluid core, from which fields emerge, rotates slower than the overlying mantle.



TERRESTRIAL DYNAMO operates in the fluid outer part of the earth's core. The nature of the core has been deduced from the paths of seismic waves spreading out from the foci of earthquakes. The outer core is known to be fluid (that is, to have a negligible resistance to shearing compared with its high resistance to compression) because it transmits pressure waves but not shear waves. The outer core is mainly composed of iron that segregated from other elements in the interior of the earth soon after its formation. Although iron itself can be a magnet, at temperatures as high as those in the core ferromagnetic materials lose their magnetic properties. The magnetism of the core is due to the electric currents flowing through it rather than to the ferromagnetism of the iron. Liquid iron has a conductivity not much less than that of solid iron, and so the core is able to carry the electric currents associated with the earth's magnetic field.

near the axis is carried around faster than those parts lying farther from the axis. The nonuniform rotation stretches the north-south field lines in an east-west direction. The resulting field is directed toward the east in the Northern Hemisphere and toward the west in the Southern Hemisphere.

As the field lines are carried around, the azimuthal field gets steadily stronger. The amplification continues until it is balanced by an opposing factor such as the tension of the magnetic lines of force or the resistive decay of the associated electric current. Edward C. Bullard of the University of Cambridge, recognizing the importance of Elsasser's work, took up the problem and showed that the azimuthal field in the core may be hundreds of times stronger than the dipole field observed at the surface. The dipole field is about .3 gauss on the surface at the Equator, but the azimuthal field in the core may be 100 gauss or more.

The dipole field at the surface of the earth must then be a secondary effect of the azimuthal field. The idea was that the upwelling of fluid in convective cells in the core deforms the lines of force of the azimuthal field, giving rise to loops of field in the meridional planes of the core that represent a net dipole field. The mechanism seemed promising, but demonstrating that convective motion would generate a dipole field proved to be a formidable task.

The problem lay in devising a mathematical representation of the velocity field and the magnetic field that dealt efficiently with the endless interactions of the two, and in devising an adequate model of the fluid motions. Bullard developed a way of representing the fields that facilitated the calculations. He then introduced the fluid velocities one might expect from a few steady, approximately circular convective cells. His calculations showed that the convection produced a variety of small-scale magnetic fields with which the convection interacted in turn, producing fields on even smaller scales. He pursued the computations to progressively smaller scales without finding conclusive evidence for the generation of a net, large-scale dipole field.

In the 1930's T. G. Cowling of the University of Leeds, who had been studying the interaction of a conducting fluid and a magnetic field, had proved a theorem stating that fluid motion cannot generate a perfect dipole field, or any field with rotational symmetry about an axis. Being aware of Cowling's theorem, Bullard deliberately placed his convective cells irregularly around the core on the hypothesis that the roughly axisymmetric field of the earth might be the residue of an asymmetric field in the core generated by these motions. Given the negative results of Bullard's calculations, it looked very much as though there might be a "super–Cowling theorem" to the effect that no magnetic field can be generated by *steady* fluid motion, whether the motion is accompanied by asymmetric fields or not. It was a black day and a dark "theorem" for those interested in the origin of the magnetic field of the earth.

I had the good fortune to work with Elsasser for a couple of years around 1954. Bullard and Elsasser corresponded frequently, and on one occasion Bullard came for a visit of several days; I could not help becoming interested in the question of the origin of the earth's field. Bullard's conjecture that there was a "super–Cowling theorem" made a strong impression on me. As a result I began to wonder what effect *nonsteady* fluid motions would have on the magnetic fields. What if the upward and downward motions in the fluid were switched on strongly but briefly and then were switched off for an extended period? For one thing rapid, short-lived fluid motions would simplify the calculations. When the motion was switched off, the electrical resistance of the fluid would cause the field to "relax" and smooth out; the complicated small-scale fields would disappear, leaving only the large-scale residue and the strong azimuthal field.

When I began to consider intermittent convection, it immediately became apparent that the more cyclonic the convection was, the more effective it would be in generating the dipole field. Cyclonic fluid motion raises and rotates the lines of force of the azimuthal field, deforming them into helixes. In the Northern Hemisphere, where the field runs from west to east, a rising cell creates a loop of field whose outer part is directed northward and whose inner part is directed southward. In the Southern Hemisphere the azimuthal field has the opposite direction, but the rotation of the cyclonic cells is also reversed, so that the helix has the same direction of rotation. Each loop in a helix has a small meridional component. When the loops diffuse and merge with neighboring loops during the time the fluid motion is switched off, they create a general meridional circulation of field that is recognized as the dipole field. Thus the intermittent cyclonic convection generates a net dipole field.

In 1958 Arvid Herzenberg of the University of Manchester proved that a magnetic field can be generated by steady fluid motions as well as by intermittent motions. Hence Bullard's conjecture proved to be wrong in the end, although it served the purpose of forcing inquiry in a fruitful direction. He had encountered a practical difficulty rather than a fundamental one: the fluid motions he chose had a relatively weak dynamo effect and the corresponding dynamo equations were correspondingly difficult to solve. It has since been shown that essentially all fluid motions have at least some small ability to generate a magnetic field. Ironically the exceptions are the simple motions that are symmetric about a plane or an axis, the

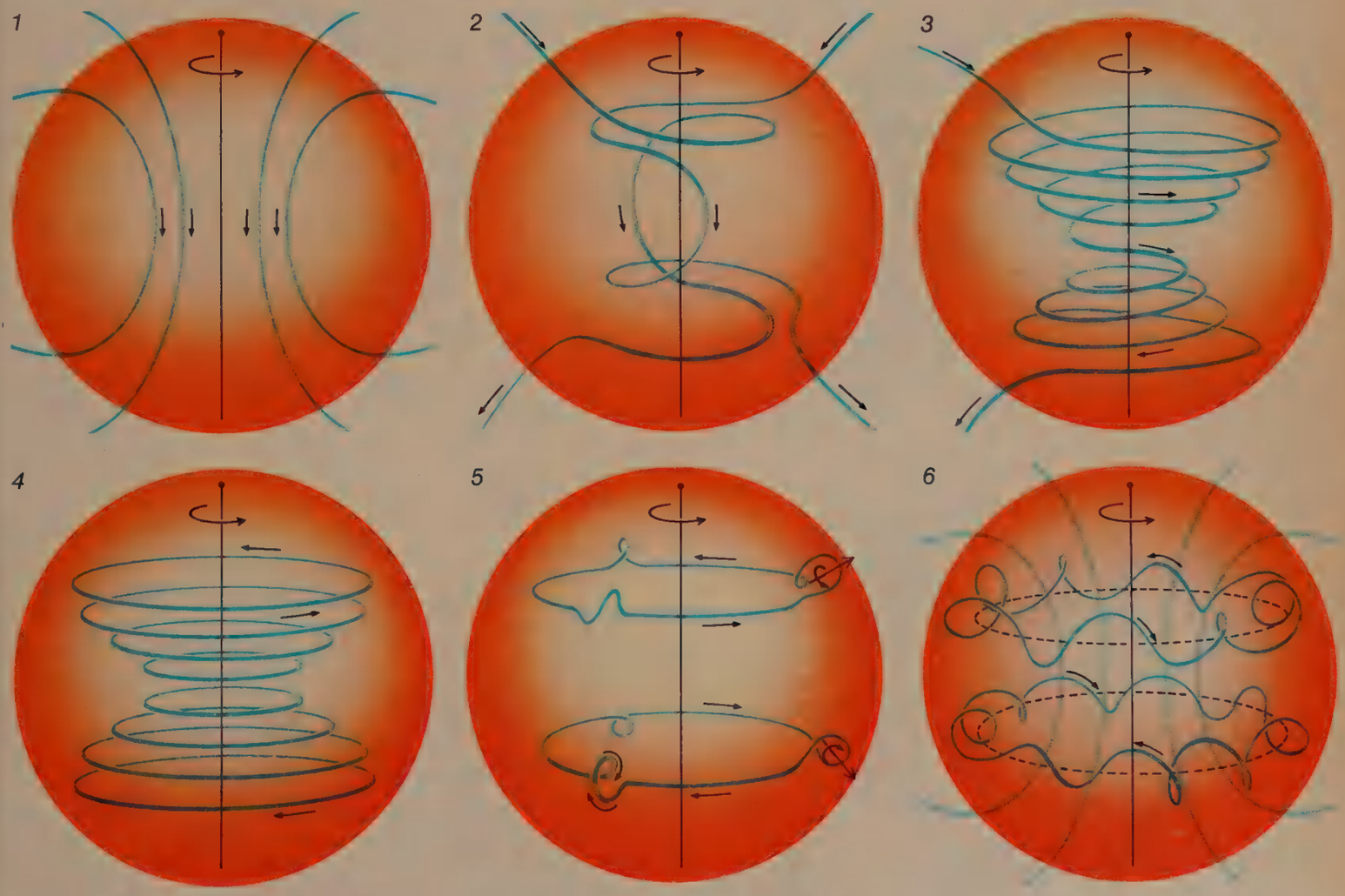
motions one would choose in seeking the simplest mathematical solutions to the dynamo equations. The essential ingredient for the generation of a field is that the motion of the fluid be helical: the fluid must rotate about its direction of motion as it streams along. Thus cyclonic convection even in the absence of nonuniform rotation is sufficient to generate a magnetic field. The naturally occurring combination of cyclonic convection and nonuniform rotation is, however, about the most efficient scheme that can be devised.

The test of the model of the terrestrial dynamo was whether or not it could be shown to amplify the dipole field at a rate high enough to balance the decay of the field. The magnetic field in a current-carrying body decays in a characteristic time (the resistive magnetic relaxation time) that is proportional to the current-carrying capacity of the body. The capacity in turn is proportional to the conductivity multiplied by the cross-sectional area. Because of its size the earth's core has a long magnetic relaxation time. If the generation of the field ceased today, it would be 30,000 years before the electric current and the magnetic field strengths fell to half their present values.

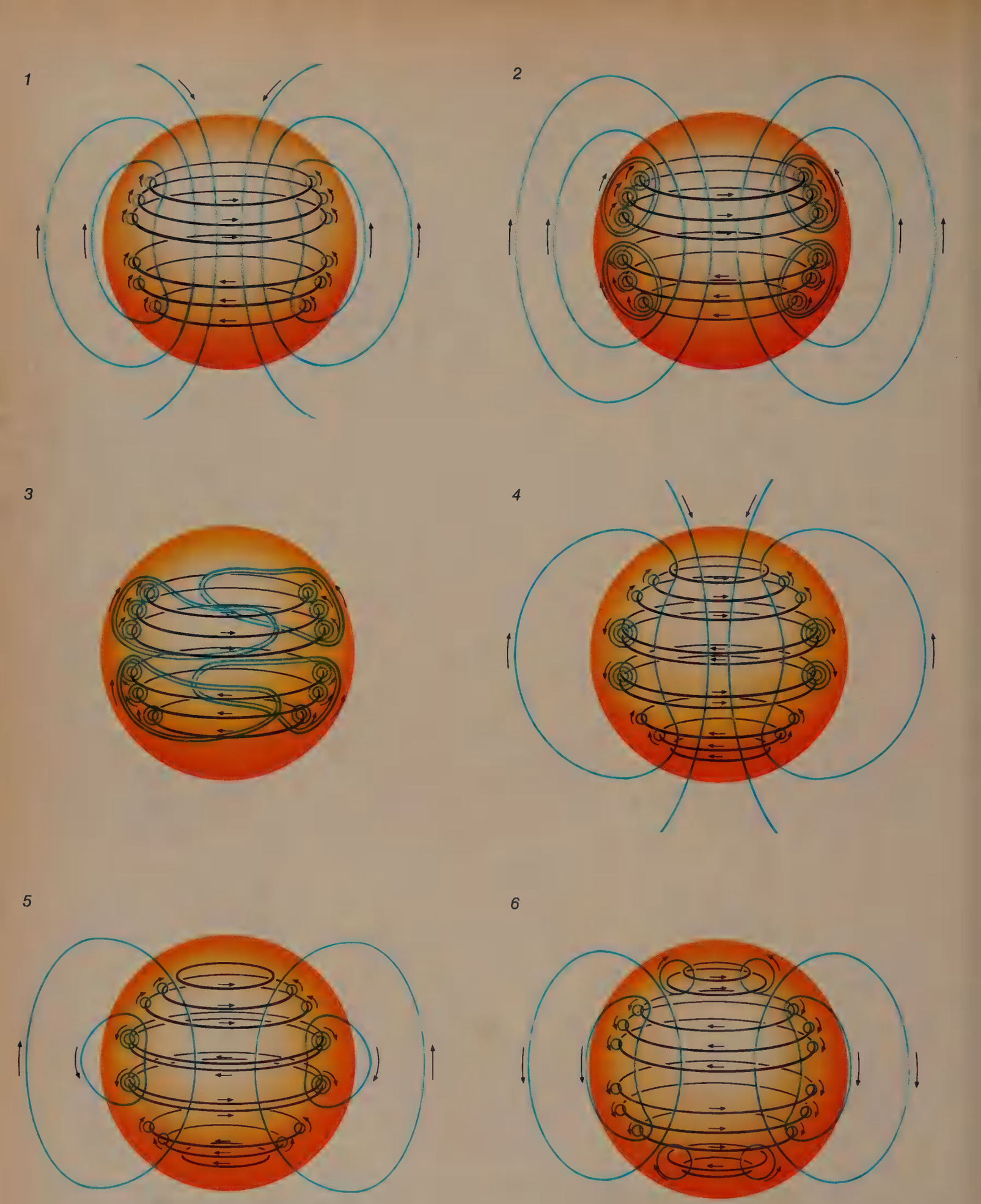
The strength of the azimuthal field is determined by the number of times the field lines are wrapped around the earth in their 30,000-year lifetime. Therefore an azimuthal field of 100 gauss or more in the core at the Equator may be associated with the dipole field of .3 gauss at the Equator. The rate at which the dipole field is generated from the azimuthal field can then be calculated on the basis of a convective velocity of one meter per hour at the surface of the core. The model of the terrestrial dynamo based on the known constraints (the strength of the dipole field, the resistivity of molten iron and the fluid velocity at the surface of the core) easily regenerates the field at a rate equivalent to replacing the dipole field once every 30,000 years, that is, at a rate equal to the rate of decay.

Another property of the earth's field that stands in need of explanation is its ability to reverse polarity. In 1955 S. K. Runcorn of the University of Newcastle upon Tyne pointed out that the alignment of grains of magnetic iron oxide in lavas indicates the direction of the local magnetic field at the time of their crystallization. He went on to show that this paleomagnetic record indicates the earth's field has reversed polarity abruptly (in only about 1,000 years) at random intervals of from 100,000 to 10 million years. It has been shown since then that a sudden change in either the location or the strength of the convective cells can throw the dynamo into a chaotic state in which the local fields run in the wrong direction and generate a large-scale field of the opposite polarity. The required changes are modest (as little as 20 degrees of latitude) and therefore plausible.

In spite of the success of the hydro-



DYNAMO MODEL of the generation of the earth's field holds that the field lines are stretched by fluid motions in the core; as a result the field is amplified at a rate high enough to balance the rate of decay. A much simplified version of the process is shown here. The north-south lines of force of the dipole field (*1*) are drawn out in the east-west direction (*2*) by the nonuniformly rotating fluid in the core. The angular velocity of the fluid decreases with distance from the spin axis of the earth, and the part of a field line in the inner core is carried around faster than the part near the core surface. The resulting azimuthal field is directed eastward in the Northern Hemisphere and westward in the Southern Hemisphere. Over their 30,000-year lifetime the field lines are wrapped around the core many times and the azimuthal field gets strong (*3*). If the original dipole field is subtracted from the spiraling field created by nonuniform rotation, the result is a purely azimuthal field (*4*). Upwelling fluid in cyclonic convection cells creates loops in the rings of azimuthal field (*5*). Rising fluid spirals counterclockwise in the Northern Hemisphere and clockwise in the Southern Hemisphere. As a result the outer part of a loop in either hemisphere is directed north and the inner part is directed south. When the loops in the helixes into which the azimuthal field lines are deformed diffuse and merge, they create the large-scale meridional field detected at the earth's surface as a dipole field (*6*).



REVERSAL OF THE EARTH'S MAGNETIC FIELD might be caused by a sudden increase in the velocity of convection in the earth's core. Ordinarily the small meridional loops (*blue*) into which the azimuthal field (*black*) is carried by upwelling fluid have time to merge, creating the large-scale dipole field (*1*). If the velocity of convection increases (*2*), the local meridional fields do not have time to merge before the nonuniform rotation of the core begins to stretch them in the east-west direction. The new azimuthal field runs in the direction opposite to that of the existing field at low latitudes and in the same direction at high latitudes (*3*). The addition and cancellation of the fields shifts the old azimuthal field to higher latitudes and produces a reversed azimuthal field at low latitudes (*4*). Convection produces reversed meridional loops of field from the reversed azimuthal field. The general meridional field created by the diffusion of these loops crowds out the original field (*5, 6*). The scheme shown here is a model; the mechanism by which the field reverses polarity is not known.

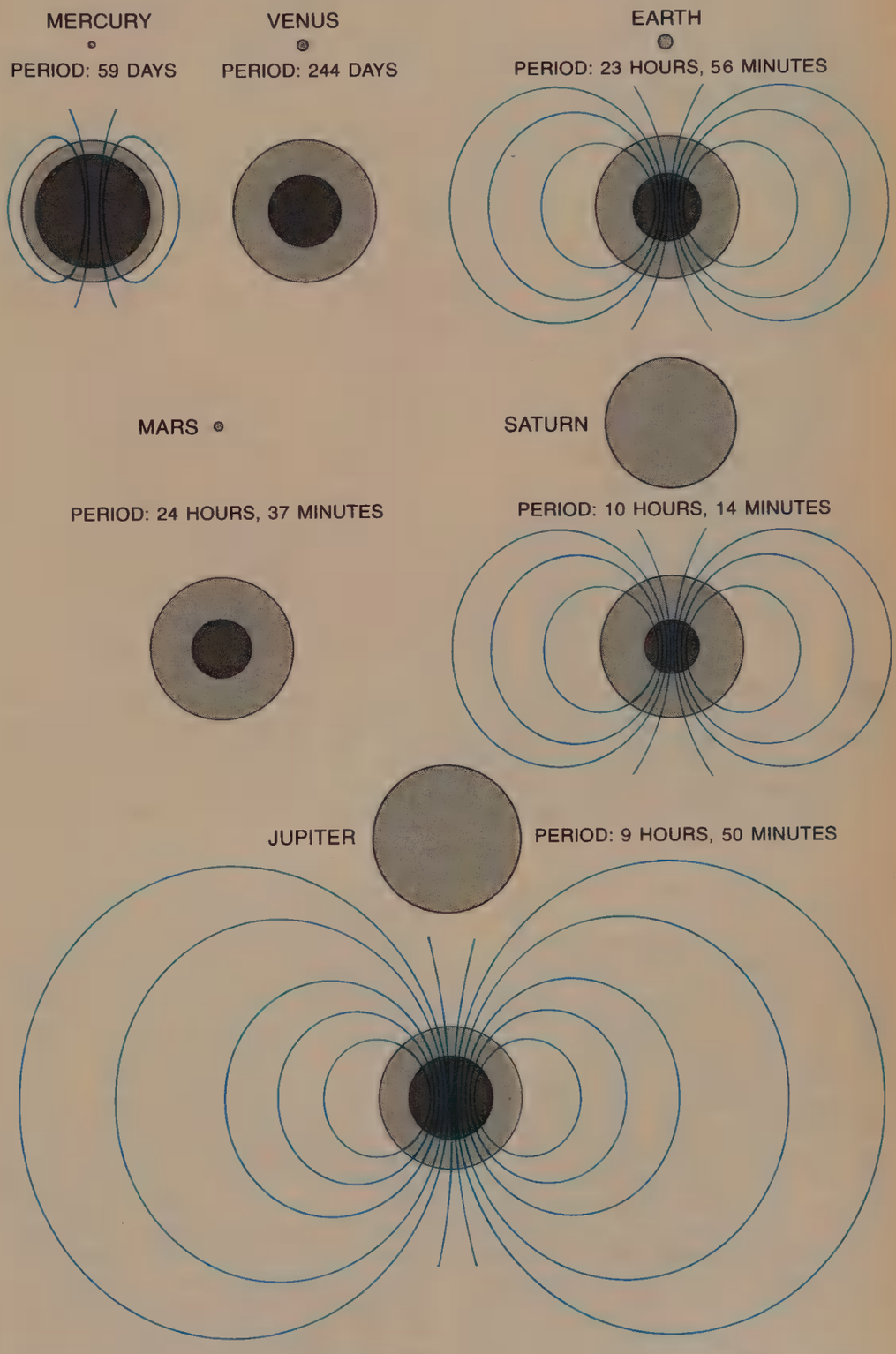
magnetic model of the terrestrial dynamo it would be misleading to suggest that the problem of the earth's field has been completely worked out. So far it has been demonstrated that fluid motions of the kind thought to exist in the core of the earth could in principle generate the observed magnetic field. The fluid motions used in the formal mathematical demonstration, however, and even those used in the sophisticated numerical experiments done recently were simplified in one or another way to make the calculations tractable. Ultimately the goal is to start from first principles, such as the rate of rotation and the energy source that drives convection. Instead of postulating a certain pattern of circulation one might deduce a precise quantitative model of the fluid motions and of the local magnetic fields in the core. The various effects of the convective flow on the field and vice versa might then be calculated.

In all probability it will be some time before the problem can be solved that completely, largely because of the lack of hard information about the source of the energy driving convection. Both the heat released by the decay of the radioactive isotope potassium 40 and the settling of a denser phase of iron through a lighter liquid phase may contribute to convection. Similarly, although sudden changes in the pattern or rate of convection could cause reversals of polarity, it is not known why there should be sudden changes. Some recent studies of coupled mathematical systems indicate that such systems have repeated erratic episodes, but it has not been shown how these results can be applied to convection in the core.

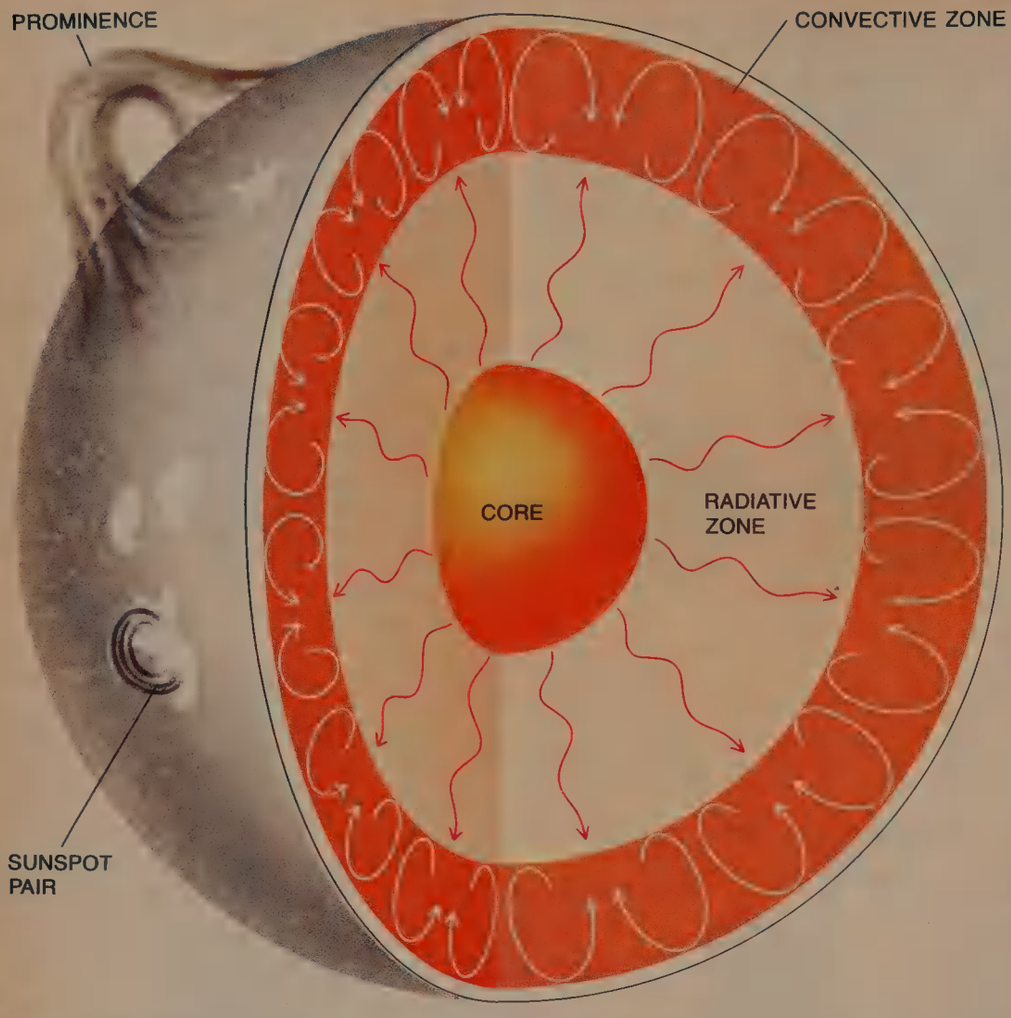
There is much less information to go on in the case of planets other than the earth. Thus the discussion of planetary fields is limited to basic questions. Why, for example, does the small planet Mercury have an intrinsic magnetic field, whereas Venus, which is roughly the size of the earth, has none? Why does Mars have no field?

The rate at which a planet generates a magnetic field is defined by what is known as the dynamo number. The dynamo number is the product of three quantities—the rotary velocity of the convective cells, the velocity gradient produced by nonuniform rotation and the volume of the core—divided by the resistivity of the body. If the dynamo number is below a threshold value, the body is incapable of generating a magnetic field as fast as the field decays. If the dynamo number is above the threshold, the rate of generation exceeds the rate of decay and the mode or spatial complexity of the growing field depends on the value of the number. For instance, a steady dipole field is generated by a body with a dynamo number lower than that of a body generating an oscillating quadrupole field.

Mercury is a small but dense planet. It has a radius about a third the radius of the earth, but its density indicates it has a relatively large metallic core with a radius about half the radius of the earth's core. On the other hand, Mercury rotates relatively slowly (with a period of 59 days). Because of the low rotation rate Mercury was not expected to generate a field. It came as a surprise, therefore, when the *Mariner 10* spacecraft registered a dipole magnetic field



STRENGTH OF A PLANET'S MAGNETIC FIELD is generally correlated with the absolute size of the planet's conductive core or layer and the planet's rate of rotation. Venus, for example, has no field because it rotates very slowly (with a period of 244 days). Both Jupiter and Saturn have strong fields because they rotate rapidly (with periods of about 10 hours) and have interior layers of metallic hydrogen and helium that are about as conductive as copper at room temperature. Saturn's surface field may be weaker than Jupiter's either because Saturn's layer of convecting metallic hydrogen is overlain by a static, perhaps rigid, layer of metallic hydrogen or because convection is weaker in Saturn than it is in Jupiter. In view of the presence of a field on Mercury the absence of a field on Mars is something of a puzzle. The size of Mars's core is not known for certain. The core may be very small or it may be nearly as large as Mercury's. Mars rotates more than 50 times faster than Mercury, yet Mars generates no field. It may be that Mars has a core too small to generate a magnetic field or that the convection in Mercury's core is unusually strong and so compensates for the small core and slow rotation.



SOLAR DYNAMO is the convective zone of the sun, which has a radius about a fourth that of the sun. This thin-shell dynamo cannot generate a meridional field fast enough to maintain a large-scale field stretching across the entire sun; the field is therefore an oscillating dipole rather than a steady one. With this qualification the operation of the solar dynamo is similar to that of the terrestrial dynamo. Twinned sunspots and solar flares are manifestations of strong local azimuthal fields. High-latitude prominences are aligned with the local meridional field lines.

as it flew by Mercury in 1974 and again in 1975.

The weak irregular fields near Venus and Mars appear to be nothing more than the sun's field carried to and pressed against these otherwise nonmagnetic spheres of rock by the wind of electrically charged particles flowing outward from the sun. Venus, often called the sister planet of the earth owing to the similarities in size and internal structure, apparently generates no field because it rotates very slowly (with a period of 244 days). Mars rotates at essentially the same rate as the earth does (with a period of 24 hours 37 minutes). The low density of Mars, however, suggests that if it has any metal core, it is a small one. The rate at which a magnetic field dissipates varies inversely with the square of the radius of the magnetic body, and so it may be that any field generated in the core of Mars is lost faster than it can be regenerated.

Comparisons among these four planets raise some interesting and as yet unanswerable questions. The field at the surface of Mercury is only a thousandth as strong as the field at the surface of the earth, but the very existence of the field is puzzling. If Mercury can maintain a steady dipole field, the earth, which rotates 59 times as fast and has a core twice as large, should be able to sustain more complicated fields. Perhaps some factor inhibits the generation of such fields in the earth's core, or perhaps fluid convection in Mercury's core is more vigorous than it is in the earth's and makes up for the smaller core and slower rotation. Mars has no field, although it rotates more than 50 times as fast as Mercury. If Mars's core is comparable in size to Mercury's, as some workers have argued on the basis of the mean density of the planet, the absence of a field on Mars and the presence of one on Mercury is baffling. Evidently Mars's core is smaller or less convective than Mercury's, and it may be much smaller and less convective. Given the meagerness of the information about the convection in planetary cores, questions such as these are likely to go unanswered for some time to come.

The sun's dynamo is a hollow spherical shell rather than a full sphere like the dynamos of the planets. The dynamo is the sun's convective zone, which begins just below the visible surface and extends a fourth of the distance to the center of the sun. The sun's period of rotation is from 25 to 35 days. The rates of rotation of solar features such as sunspots show that the surface of the convective zone rotates nonuniformly, the surface at the poles having an angular velocity about two-thirds the velocity at the equator. The convection is visible as granulation, the rice-grain markings seen in photographs of the sun's surface. Because the sun is rotating, the convection is undoubtedly cyclonic.

The solar dynamo operates much like the terrestrial one. The sun's azimuthal field, unlike that of the earth, is directly observable because the top of the convective zone can be seen. Twinned sunspots, and bipolar magnetic regions in general, are local bulges in the azimuthal field pushed to the surface of the sun by magnetic buoyancy. The azimuthal field below the surface is at least as strong as 200 gauss and may be as strong as 10,000 gauss. The meridional field generated from the azimuthal field is visible only at high latitudes, where it shows up in coronal streamers and polar prominences. It is much weaker than the azimuthal field, having a strength of perhaps five to 10 gauss.

The sun has a magnetic cycle with a period of 22 years. The coordinated waxing and waning of the azimuthal and meridional fields in the course of the cycle cause the field to reverse polarity about once every 11 years. It can be shown that this magnetic dance is a consequence of the characteristics of the solar dynamo. A thin-shell dynamo such as the sun's loses field too rapidly for the generated field to reach diffusive equilibrium over the full length (in latitude) of the shell. The low-latitude field has not had time to reach equilibrium before an opposite field is generated at high latitudes. Hence a thin-shell dynamo has no steady dipole mode; the lowest available self-sustaining mode is an oscillating dipole.

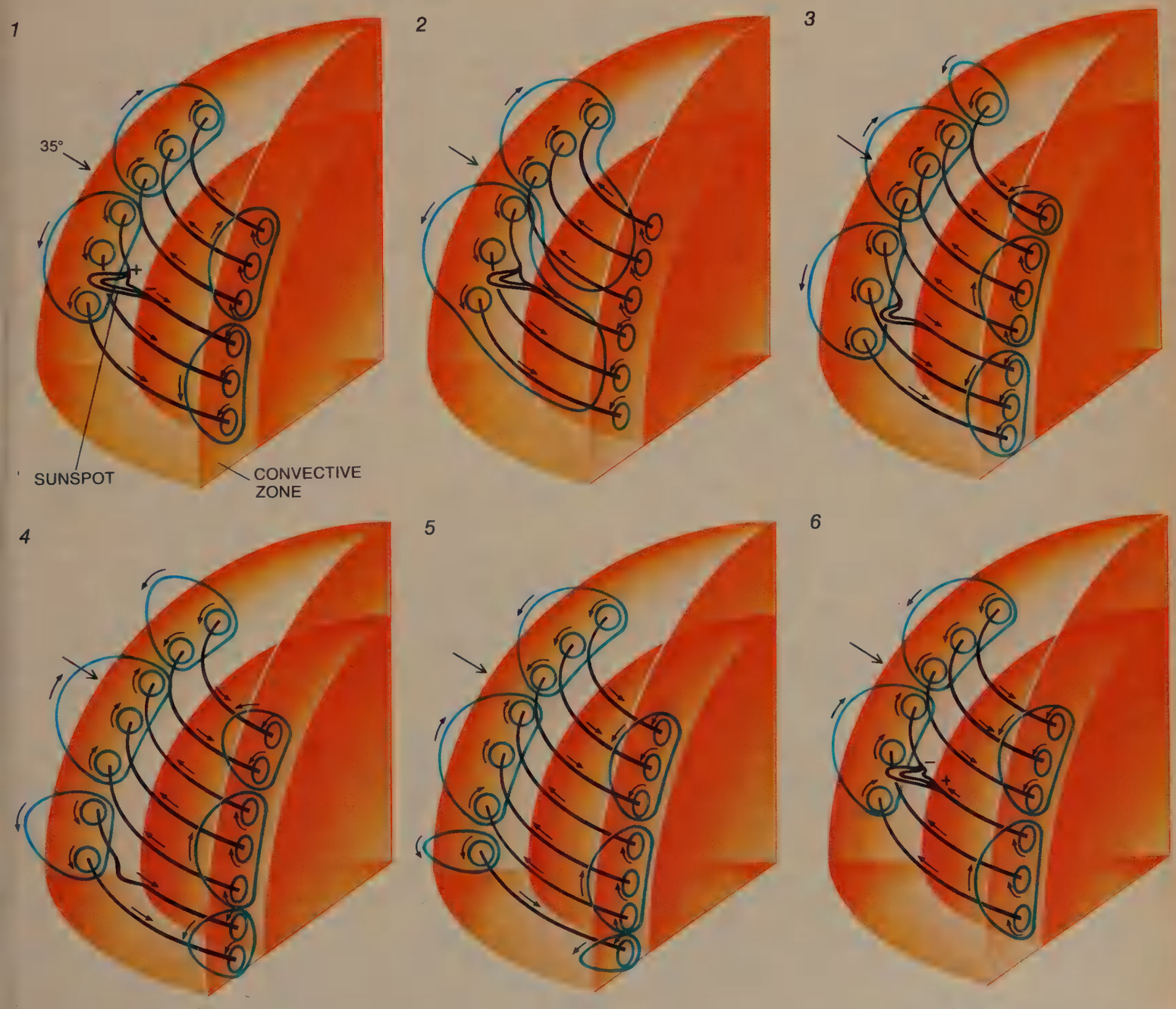
The dynamo mechanism as the explanation for the fields of some of the magnetic stars and of the galaxies does remain challenged by the rival hypothesis that the fields are primordial, trapped in the stars and galaxies since the formation of the universe. Stars are formed by the gravitational collapse of interstellar gas. The interstellar gas is threaded by the galactic magnetic field, so that a volume of gas carries some of the field with it when it collapses to form a star, compressing the field by the same factor as the surface area of the gas is compressed. Much of the trapped primordial field escapes from the star in the course of the vigorous mixing that accompanies the final stages of collapse, but it is not unreasonable to suppose some significant amount is retained.

The planets were formed in a similar

way, but in their case there is a simple argument against primordial fields. As I have mentioned, the magnetic relaxation time of the earth's core is about 30,000 years, far shorter than the earth's lifetime. The earth was formed more or less at the same time as the sun, some 4.5 billion years ago. Left to itself a primordial field would have disappeared long ago; the dynamo mechanism must be invoked to explain the presence of a field today. In the case of the stars the argument is less conclusive. As Cowling pointed out more than 30 years ago, the resistive-decay time of a star such as the sun is typically five billion years, comparable to the age of the sun. The sun's magnetic cycle can only be explained on the basis of a dynamo, but the possibility that there is also an undetected primordial field as strong as a million gauss trapped in the stable core of the sun cannot be excluded.

The hypothesis of primordial fields is more interesting in the case of stars with fields much stronger than the sun's. One class of stars with strong fields consists of the magnetic *A* stars. (*A* stars are stars with a surface temperature of about 10,000 degrees Kelvin.) They were discovered in the late 1940's by Harold D. Babcock and Horace W. Babcock of the Mount Wilson and Palomar Observatories, who measured their fields with a sensitive magnetometer they had developed. The fields of these stars range in strength from a few hundred gauss to 34,000 gauss.

It has since been shown that the field of a magnetic *A* star is a dipole that rotates more or less rigidly with the star itself. Moreover, the axis of the field tends to be perpendicular to the axis of rotation instead of approximately parallel as it is in planets and other stars. In addition the stars have a modest rate of rotation (a period of a month or so), and the more rapidly rotating stars tend to have weaker fields than the slower stars do. The lack of correlation between the strength and the mode of the field, be-



SUN'S MAGNETIC CYCLE begins with two bands of azimuthal field of opposite sign in each hemisphere (*1*). The band at high latitudes shows up as the "ephemeral" bipolar magnetic regions visible in magnetograms of the sun. The lower band shows up as twinned sunspots, the entry and exit points of an azimuthal field bulging above the surface of the convective zone. The meridional field generated from each band of azimuthal field by convection is stretched by the nonuniformly rotating gas of the convective zone into new azimuthal field (*2*). The new field reinforces the original field at the southern edge of the bands and cancels it at the northern edge (*3*). As a result the bands of azimuthal field migrate toward the equator and a new band of opposite sign is created at high latitudes. At the equator the bands of azimuthal field are canceled by bands of opposite sign arriving from the other hemisphere (*4*). The sunspots associated with each band of field decrease in number and strength as the band slowly decays (*5*). The recovery from sunspot minimum to a new maximum is relatively rapid (*6*). This sequence of drawings shows one reversal of polarity, or one-half of the roughly 22-year solar magnetic cycle.

tween the strength of the field and the rate of rotation and between the orientation of the field and the axis of rotation makes it unlikely that the fields of these stars are generated by the dynamo mechanism.

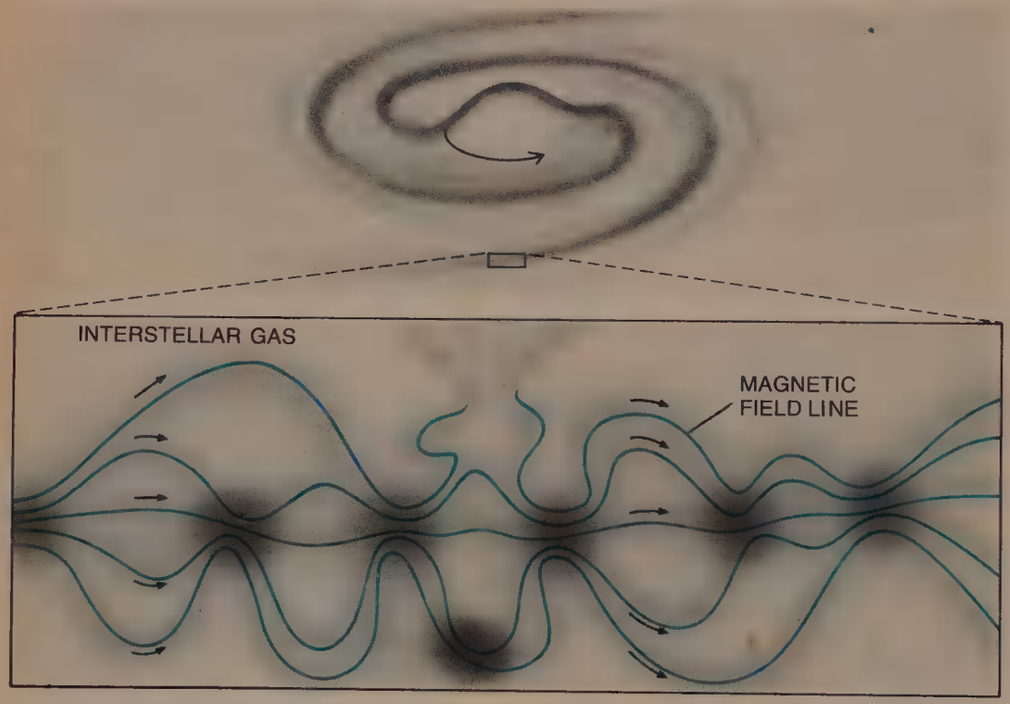
It has also been suggested that the galactic magnetic field is primordial because the resistive-decay time of a body as large and as conductive as the galaxy is longer than the galaxy is old. The effect of the electrons circulating around the galactic field lines in interstellar space on the plane-polarized radio waves from distant sources provides a measure of the field strength. Observations of the variation of the plane of polarization with frequency (the Faraday rotation effect) show that the field is between 2×10^{-6} and 3×10^{-6} gauss. Presumably all that would be needed to account for the present field is that the diffuse gas from which the galaxy formed was threaded with a field of about 10^{-9} gauss. This field would be compressed when the gas collapsed to form the galaxy. It has been argued that the compression and subsequent stretching of the lines of force in the nonuniformly rotating galaxy could have produced a field of the observed strength. Recent work by Jacques P. Vallée of the National Research Council of Canada, however, places an upper limit of 10^{-11} gauss on the present intergalactic field, a field strength much lower than the 10^{-9} gauss generally assumed in arguments for the primordial origin of the galactic field.

What is more, the system composed of the interstellar gas, the cosmic rays and the galactic magnetic field is highly unstable. The gas tends to gravitate toward the central plane of the galactic disk and the magnetic field tends to billow upward, away from the plane. Although there is not enough information to build a quantitative model of the turbulent loss of field, such calculations as are possible indicate that the vigorous small-scale undulations of the field cause it to wiggle out of the gas in the disk at a rate fast enough to preclude the possibility that a primordial field has persisted to the present day. Turbulent loss reduces the decay time of the field to about 300 million years, whereas the galaxy is 10 billion years old.

The dynamo mechanism successfully explains what is known about the galactic field. The dynamo equations recast in the appropriate form for a thin disk with the effective conductivity of the turbulent interstellar gas show the galaxy could regenerate a magnetic field at a rate that would balance the loss by turbulent escape and resistive decay. The gaseous disk of the galaxy is known to rotate with an angular velocity that decreases outward from the center. Supernova explosions and the sudden formation and lighting up of massive luminous stars produce violent turbulence or convection. Because of the rotation, the turbulence is cyclonic. The primary field of the galaxy is azimuthal, extending around the disk of the galaxy. The meridional components of the galactic field cannot be observed because they are lost in the strong local fluctuations of the azimuthal field.

One last question deserves consideration even though it is, by virtue of the lack of evidence, more a philosophical question than a scientific one. The dynamo mechanism does not provide for the outright creation of a magnetic field but only for the reproduction, that is, the amplification, of existing fields. Where then did the fields the dynamos amplify come from? It may not be necessary to resort to the idea of a primordial field, the existence of which would in turn need to be explained. In 1941 Ludwig F. Biermann of the Humboldt University of Berlin pointed out that there are "atomic battery" effects in the moving ionized gas of the stars. For example, the isotherms, or surfaces of equal temperature, in a rotating star do not coincide with the oblate level surfaces at which the gravitational and centrifugal potentials are balanced. As a result of the unbalanced forces free electrons slide along these surfaces, creating weak meridional currents. If a star was initially free of magnetic field, the Biermann battery effect would over an extended period build up a weak azimuthal field. The battery effect alone cannot account for fields of the strength observed today, but it does guarantee that if all else fails, the stars and galaxies are seeded with weak magnetic fields on which the powerful hydromagnetic mechanism can then operate. Similar effects due to slight radioactivity or to chemical or thermal inhomogeneity would seed the planets with weak fields.

When all the evidence is considered, it looks as though magnetic fields are generated by fluid motions on all scales from the 2,000 kilometers of the liquid-metal core of a planet to the million kilometers of a star to the 10^{18} kilometers of a galaxy. Much remains to be learned both about the dynamos and about the unusual bodies that have fields too strong to be explained by the known dynamo effects, but the existence of these tantalizing unsolved problems is what makes the subject as interesting as it is.



GALACTIC MAGNETIC FIELD is predominantly azimuthal, that is, parallel to the plane of the galaxy, although the direction of the field fluctuates widely. The field's direction was originally determined from the polarization of light from distant stars. Dust grains in the interstellar gas are set spinning by collisions both with one another and with hydrogen atoms. Each dust grain carries an electric charge and hence tends to line up with its long axis perpendicular to a magnetic field line around which it spins. The polarization of the light absorbed and reemitted by the dust lying between a star and the solar system is therefore determined by the orientation of the galactic magnetic field. The strong local fluctuations of the magnetic field are caused by its interaction with the interstellar gas and with the cosmic rays it confines and by which it is inflated. It is thought supernova explosions of massive stars blow the interstellar gas and cosmic rays outward from the plane of the galaxy, creating the spherical shells of gas visible on maps of the radio emission of atomic hydrogen, one component of the gas. The magnetic field threaded through the gas is carried with it and compressed in the periphery of the shell. If the magnetic field is not strong enough to balance the outward pressure of the gas, the gas and the entrained field are swept off into space. Such turbulence contributes significantly to the rate at which the galactic field dissipates. The rate of loss makes it unlikely the field is primordial and supports the argument that the field is regenerated by rotating and convecting gas.

Museum piece, circa 1987.

The clapstick may be taking an early retirement. This and other conventional tools of the filmmaking trade may find themselves relegated to crates and carted off to museums.

What will replace them? Kodak film with Datakode magnetic control surface! This radical advance in film manufacture, which unites chemical and magnetic imaging technologies, gives film the ability to "converse" with computers. This achievement could substantially reduce the time and costs associated with film postproduction.

The Datakode magnetic control surface is a thin, transparent layer of iron oxide particles (approximately 9 billion per square inch) coated across the entire back of the film during manufacture. Less than 8 microns thick, this transparent layer provides the means to record machine-readable information (100 bits of binary data in a single track on each frame of film), and makes possible a uniform frame-indexing code that can be used with both film and videotape. Best of all, it accomplishes this interface without altering the quality or characteristics of the final image.



In the not-too-distant future, the use of discs, video displays, time-code synchronization, and automated printing equipment will speed filmmakers through all the noncreative, repetitive steps associated with film postproduction.

The new Datakode magnetic surface is more than just another milestone for Kodak and a boon to filmmakers. It represents the most significant development in motion picture technology in the last 50 years. It's a victory for science. And for the chemists and researchers who met the challenge to add to film's capabilities.

For more information on the new Datakode surface, send your request to: Eastman Kodak Company, Department GBSA-7, 343 State Street, Rochester, NY 14650.

Kodak

Kodak. Where technology anticipates need.



© Eastman Kodak Company, 1983


Only in


a Jeep.®
Jeep® CJ
SAFETY IS A SNAP

SCIENCE AND THE CITIZEN

From MIRV to MHV

Perhaps the most poignant aspect of the report by the President's Commission on Strategic Forces (the Scowcroft commission) is the implicit acknowledgment that in retrospect the U.S. decision of a decade or so ago to begin deploying multiple independently targetable reentry vehicles (MIRV's) on the nation's land-based strategic missiles was a mistake. The destabilizing effect on the strategic balance of highly accurate, "counterforce-capable" multiwarhead missiles has come to be generally recognized by U.S. military planners only after the U.S.S.R. followed the American lead and installed MIRV's on its long-range missiles, thereby threatening the land-based component of the U.S. deterrent forces. The response of the Scowcroft group and others is to advise that the U.S. now plan to revert ultimately to a single-warhead missile (nicknamed Midgetman) and try to persuade the Russians to do the same. In the words of former national-security adviser McGeorge Bundy, the introduction of MIRV technology ranks as "the United States' worst single contribution to the nuclear arms race."

The mistake is about to be repeated, in the view of many long-term students of the arms race. This time, however, the major technological initiative in question is the imminent introduction of a new generation of antisatellite, or ASAT, weapons. According to a petition prepared early this year by Richard L. Garwin and Carl Sagan and signed by more than 40 prominent American physicists, space scientists and defense experts, "the testing or deployment of any weapons in space [including antisatellite weapons] significantly increases the likelihood of warfare on earth." The petition, sent to the leaders of "all spacefaring nations with an independent launch capability," calls for the prompt adoption of a treaty prohibiting the extension of the arms race to space, noting that "if space weapons are ever to be banned, this may be close to the last moment in which it can be done."

At present neither the U.S. nor the U.S.S.R. has a reliable ASAT system. For the past 15 years or so the Russians have apparently been testing a rather primitive ASAT weapon consisting of a nonnuclear explosive device launched into low earth orbit by an SS-9 intercontinental ballistic missile (ICBM). The system has reportedly succeeded in destroying an orbiting target satellite in about half of some 20 attempts. Most U.S. military satellites, including those responsible for long-range communications and early warning of a strategic-missile attack, are in very high, geosynchronous or semisynchronous orbits and hence are considered invulnerable to the current Russian ASAT system.

Meanwhile the U.S. has developed a quite different ASAT weapon. It consists of a small, nonexplosive interceptor called the Miniature Homing Vehicle (MHV), which is mounted on a two-stage rocket designed to be launched into space from a high-flying F-15 fighter plane. The MHV is equipped with a combination of infrared telescopes, a laser gyroscope and an array of small jets that will enable it to seek its target and destroy it by direct impact at very high velocity. In its present form the MHV cannot reach satellites in geosynchronous orbits; unlike the U.S., however, the U.S.S.R. currently has almost all its "space assets" in either low or highly elliptical orbits, where they are considered more vulnerable. With a three-stage booster the U.S. ASAT weapon should be able to reach geosynchronous orbits, a capability the Russians could attain with their present system only by recourse to a new, much larger ground-based rocket.

The first test launch of the U.S. interceptor is expected soon, with tests against space targets scheduled for later this year or early next year. It is estimated that the MHV system could be operational as early as 1987. According to the Air Force, the projected total cost of the ASAT program is $3.6 billion; unofficial estimates of the ultimate cost of the system by the General Accounting Office run as high as "tens of billions."

Bilateral negotiations to forestall the impending arms race in space are currently in abeyance, having been suspended by the Carter Administration in 1979 after the Russian intervention in Afghanistan. In 1981 the U.S.S.R. submitted to the United Nations a draft treaty calling for a "prohibition on the stationing of weapons of any kind in outer space" but making no reference to ground-based systems. Recently Foreign Minister Andrei A. Gromyko appeared to broaden the terms of the Russian proposal, offering to ban "in general the use of force both in space and from space against the earth." So far the Reagan Administration has not responded officially to these initiatives, in spite of several resolutions introduced in the Senate and the House of Representatives calling for renewed negotiations and for a moratorium on the flight testing of weapons in space.

In the absence of any apparent interest on the part of the administration in a negotiated ASAT agreement a private group, formed in March under the auspices of the Union of Concerned Scientists, has drafted a "model treaty" intended "to focus attention on the issues that such negotiations are likely to encounter." Unlike the Russian draft treaty, the treaty proposed by the U.S. scientists would prohibit the testing of all ASAT systems, including ground-based ones, against objects in space.

Testifying before the Senate Foreign Relations Committee in May, Kurt Gottfried of Cornell University, chairman of the scientists' group, stated that he and his colleagues "have compared the status and potential capabilities of both ASAT systems and conclude that an immediate moratorium on ASAT space tests would not jeopardize our security if negotiations were to fail, or if a treaty were eventually abrogated or circumvented by the Soviet Union." Noting the "striking parallel" between the present competition in ASAT weapons and the history of MIRV technology, Gottfried concluded that the lesson of the MIRV episode "applies directly to antisatellite weapons. The Soviets have been both foolish and reckless to spend some 15 years nurturing a clumsy threat against a rather small portion of our satellites. Their major accomplishment has been to provoke us into building a far more sophisticated system. Our ASAT, if deployed, will give us a temporary advantage. But as with ballistic missiles, an ongoing competition in space weaponry will, inexorably, reduce the security of both sides."

The Basalts of Venus

A milestone in the exploration of the cloud-shrouded surface of Venus came in March of last year, when two Russian spacecraft landed on the surface, sampled Venusian rock and determined its chemical composition. *Venera 13* landed on an "upland rolling plain," *Venera 14* landed on "lowland." (According to radar mapping by the *Pioneer* Venus probe, these terrains make up 92 percent of the surface.) Writing in *Journal of Geophysical Research,* a group led by Yu. A. Surkov of the V. I. Vernadsky Institute of Geochemistry and Analytical Chemistry in Moscow now gives details of what the spacecraft found.

Each spacecraft carried "a miniature drilling rig" that took a sample of the surface. The sample was conveyed through an "air lock" to a "measuring cell." There it was bathed in radioactivity emitted by plutonium 238 and iron 55. The former isotope elicited the emission of X rays at characteristic frequencies from magnesium, aluminum and silicon atoms; the latter elicited the emission of X rays from potassium, calcium and titanium. The X rays were

analyzed by on-board spectrometers. Some initial calibration spectra were made while the measuring cell was empty. Then, 224 seconds after each landing, Venusian soil arrived in the cell. *Venera 13* succeeded in measuring 38 spectra; *Venera 14* measured 20. The investigators compared the spectra with those of 200 terrestrial and lunar "rocks of well-known elemental composition."

The rock of the rolling upland proved to be "close to the composition of potassium alkaline basalts of the earth's crust," an uncommon type of volcanic rock found in islands and along rift zones in the Mediterranean. Such zones mark the divergence of crustal plates as new ocean basins form. The Venusian rolling upland is thought to be quite different. It is high and smooth and is pocked by myriad craters; thus it is taken to "represent the ancient crust of Venus." The rock of the lowland proved to be similar to "basalts of the oceanic crust of the earth." The Russian investigators surmise "that Venusian lowlands are covered by basaltic lava flows."

The investigators offer some "general ideas that arise from looking at the data." They note that the earth, Venus and the moon have two main types of terrain, "ancient crust (continents on the earth, rolling uplands on Venus and highlands on the moon) and younger formations (the oceanic crust on the earth, lowlands on Venus and maria on the moon)." Nevertheless, Venus differs notably from the earth and the moon. The rocks are not as diverse. "Perhaps a considerable part of the surface of Venus is covered by basaltic rocks that have originated from different depths and at different times of the formation of the crust."

Doubtful Diets

Many physicians maintain that the health of Americans would be markedly improved if only people would modify a few simple behaviors rather than depending on new kinds of medical intervention. The trouble is that people need information on what behavior to modify and how, and (except in the case of cigarette smoking) the signals from public-health experts have been uncertain. The prevention or correction of high blood pressure by dietary change is a case in point. From 35 million to 60 million Americans (depending on who is making the estimate) have hypertension, and perhaps 95 percent of them have "essential" hypertension: elevated arterial blood pressure that cannot be attributed to any identifiable disease or condition. For many years correlations have been reported between blood pressure and dietary factors ranging from an excessive intake of calories (obesity) to a deficiency of magnesium, in some cases suggesting cause-



The 757 is the most fuel-efficient jetliner in the sky. In passenger comfort, it's superior to any other plane its size. Here's the perfect match of advanced technology and passenger conveniences. It's one way Boeing is helping to keep air fares one of the world's best travel values.

BOEING
Getting people together.

Speak French like a diplomat!

What sort of people need to learn a foreign language as quickly and effectively as possible? *Foreign service personnel,* that's who.

Now you can learn to speak French just as these diplomatic personnel do — with the Foreign Service Institute's Basic French Course.

The U.S. Department of State has spent thousands of dollars developing this course. It's by far *the most effective way* to learn French at your own convenience and at your own pace.

The Basic French Course consists of a series of cassettes and an accompanying textbook. Simply follow the spoken and written instructions, listening and repeating. By the end of the course, you'll be learning and speaking entirely in French!

This course turns your cassette player into a "teaching machine." With its unique "pattern drill" learning method, you set your own pace — testing yourself, correcting errors, reinforcing accurate responses.

The FSI's Introductory Basic French Course comes in two parts, each shipped in a handsome library binder. Part A introduces the simpler forms of the language and a basic vocabulary. Part B presents more complex structures and additional vocabulary. Order either, or save 10% by ordering both:

- ☐ **Basic French, Part A.** 12 cassettes (15 hr.), and 194-p. text, $125.
- ☐ **Basic French, Part B.** 18 cassettes (25 hr.), and 290-p. text, $149.

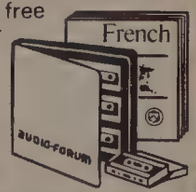
(Conn. and N.Y. residents add sales tax.)

TO ORDER BY PHONE, PLEASE CALL TOLL-FREE NUMBER: **1-800-243-1234.**

To order by mail, clip this ad and send with your name and address, and a check or money order — or charge to your credit card (AmEx, VISA, MasterCard, Diners) by enclosing card number, expiration date, and your signature.

The Foreign Service Institute's French course is unconditionally guaranteed. Try it for three weeks. If you're not convinced it's the fastest, easiest, most painless way to learn French, return it and we'll refund every penny you paid. Order today!

81 courses in 36 other languages also available. Write us for free catalog. Our 10th year



Audio-Forum
Suite 19K
On-the-Green,
Guilford, CT 06437
(203) 453-9794

AUDIO-FORUM®

Or visit our New York sales office: 145 E. 49th St., New York, N.Y. 10017 (212) 753-1783

and-effect relations and implying that a particular modification of diet might be prophylactic or therapeutic.

Such findings have now been reviewed and evaluated in 35 papers published as a supplement to *Annals of Internal Medicine.* The signals are still unclear. As Harriet P. Dustan of the University of Alabama School of Medicine puts it in an *Annals* editorial, "we are far from identifying a dietary factor important in the pathogenesis of hypertension." The idea of diet modification nonetheless has attractions. Epidemiological evidence does show differences in the diet of populations with differing prevalences of hypertension, and diet is in theory readily subject to modification by an individual on his own.

A further reason for interest in the dietary approach by some physicians is disenchantment with aggressive drug therapy for mild hypertension. Within the past 10 years or so, as new antihypertension agents have become available, it has become almost axiomatic that any degree of hypertension should be treated. For most physicians this means prescribing drugs for patients with a diastolic pressure of 90 or more. Norman M. Kaplan of the University of Texas Health Science Center at Dallas takes issue with the practice, arguing that the data do not in fact support drug therapy for asymptomatic mild (90 to 104) hypertension and that any drug treatment has some risks. Nondrug therapies, including dietary modifications, are indicated for such patients, he says.

What modifications? The doctors disagree. Consider reduction of salt (sodium chloride) intake, perhaps the commonest prescription. James C. Hunt of the Center for the Health Sciences of the University of Tennessee holds that "excessive sodium consumption by the American public is a justified cause for concern." While conceding that hypertension is more directly related to sodium intake in some people than in others, he cites cross-cultural studies showing that societies whose sodium intake is elevated have a high frequency of hypertension and of stroke, one of its major sequels. On the other hand, within a particular society many studies have failed to show a correlation of elevated sodium intake with hypertension, leading other investigators to doubt the value of sodium restriction for the general population. Dustan implies that the case for cutting back on salt is not proved. The consensus nonetheless appears to be that since reduction of sodium may prevent the development of hypertension in some people and since a high-salt diet is almost certainly not beneficial, reduced salting of food and reduced consumption of salty snack foods is probably a good idea.

The situation is clearer with regard to excessive alcohol consumption and obesity. Both are shown to contribute to hypertension, and to do so independently of each other. Dustan notes in particular that "obesity is so prevalent and so closely associated with hypertension" that more attention should be given to its control and to understanding why it causes hypertension.

Unraveling Knot

It is well known that in industrial countries the institution of marriage is in a state of change. How much has it changed and what is its future? Writing in *Bulletin of the American Academy of Arts and Sciences,* Kingsley Davis of the University of Southern California examines demographic data on marriage in the most prosperous industrial countries since World War II. He concludes that marriage is losing its connection to its two most significant functions: providing domestic intimacy and offering a framework for raising children. Further, he predicts that the relation between marriage and childbearing will become increasingly tenuous without fundamental social reforms.

From 1945 through 1960 marriage gained in the industrial countries. In general there was a decrease in the age at which men and women first married and an overall increase in the fraction of the population that was married. Beginning in the 1960's, and particularly in the 1970's, both trends were reversed. In 1960, 14 percent of all women aged 30 through 34 in the U.S. were unmarried; by 1980 the fraction was 27 percent. Other industrial nations show a similar increase in the fraction of single women. The change is due both to a rise in the divorce rate and to a tendency for young people to postpone marriage. After a steady decline since 1945 the age of first marriage began to rise in about 1970.

Changing social customs have resulted in a large group of young, single people. Unmarried people are not, however, forgoing the pleasures of cohabitation. From 1977 through 1980 the number of married couples in the U.S. changed little but the number of unmarried couples living together rose by 63 percent. In Norway in 1977 among all women who were 18 or 19 years old and were living with a man 43 percent were not married. The fraction of women living in such "consensual unions" decreases with age, but among younger women informal unions are competing vigorously with formal ones.

Some have suggested that a consensual union is simply a marriage without certain more or less trivial legal trappings, but Davis notes that there are significant differences between the two. Informal unions are made by younger people, are less stable and result in fewer children than marriage does. In any event the fraction of the population in any kind of union, legal or not, is decreasing; thus the spread of consensual unions has not compensated for the decline in the marriage rate.

Not only is marriage getting competition from other kinds of relationships in providing domestic intimacy; it is rapidly losing its monopoly on reproduction and child rearing. Married couples are becoming steadily less fertile. From 1945 through 1949 the net reproduction rate in 19 of the most economically developed countries was 1.27. (The net reproduction rate indicates whether the population is replicating itself; the replacement level is 1.) From 1975 through 1979 the rate in the 19 countries was .87, well below replacement level. In countries such as the U.S., the U.K. and Sweden the rate would have been considerably lower if it had not been for the large number of immigrants from poorer countries, who have relatively high fertility.

Furthermore, the reduction in fertility among married women is much greater than the change in the reproduction rate suggests. The reason is that it has become much commoner for an unmarried woman to have a child. In 1950 a married white woman in the U.S. was 17 times as likely to have a child as an unmarried one; in 1970 the figure had dropped to four times as likely. Among U.S. blacks an unmarried woman is currently about as likely as a married woman to have a child. The fertility of consensual unions is quite low, hence most births to single women are to women who are living alone.

If the children of divorced parents are added to children of single mothers, it can readily be seen that a substantial fraction of children in developed countries are growing up in a setting other than the traditional marriage. In 1970, 15 percent of all U.S. children under 18 did not live with both of their biological parents; in 1981 the fraction was 24 percent. Fewer children are currently being raised in intact marriages. Conversely, fewer marriages include children: in 1981, 49 percent of all U.S. married couples had no children of their own under 18 living with them.

The data accumulated by Davis show that marriage is being separated from its traditional functions. Conjugal relations are more frequent in an informal union and childbearing is more frequent outside marriage, indeed outside any kind of union. In accounting for the changes in marriage and in particular the declining fertility of married couples Davis cites several factors. One of the most significant is the increasing likelihood that a married woman will work outside the home. About 60 percent of married women in the U.S. currently have such jobs, only slightly less than the rate of 65 percent for single women. In the absence of affordable arrangements for child

care the greater the fraction of working women is, the lower the fertility rate is likely to be. According to Davis, if the accustomed functions of marriage are to be preserved in industrial societies, provisions must be made to reduce the conflict between work and motherhood. He concludes: "Although people value children very highly, it seems likely that they will not exceed replacement-level fertility unless incentives for childbearing and child care are systematically improved and disincentives alleviated, which in large part means reforming the institution of marriage. Failing such a development, it seems likely that the population of the industrial nations will increasingly be sustained by immigrants from the Third World."

Nonfarming for Farmers

What does a good farmer do with fertile land he undertakes not to farm? The question comes up because many farmers in the U.S. will be withdrawing land from commercial production this year under the Farm Acreage Reduction Program, the Federal effort to deal with chronic surpluses of certain crops and the substantial amounts of farm products in storage from previous harvests. Participating farmers will receive cash payments for withdrawing 20 percent of their acreage and payments in kind from stored or current harvests for additional withdrawals. Advice on what to do with the withdrawn land is proffered in *Crops and Soils Magazine,* a publication of the American Society of Agronomy, by D. R. Hicks, N. P. Martin and E. A. Oelke of the University of Minnesota.

The program for withdrawal covers mainly corn and wheat. Hicks and his colleagues address their advice to corn farmers, but the recommended practices would be similar for producers of other crops. The advice is based on regulations issued by the Agricultural Stabilization and Conservation Service of the United States Department of Agriculture, which is administering the withdrawal program. The regulations offer the corn farmer three options: (1) to establish annual or perennial crops of grasses or legumes or both; (2) to plant a small crop of grain that would be made unharvestable before it ripens, and (3) to leave on the ground the residue of the 1982 corn crop and control the weeds that grow this year. Two other stipulations are that the diverted acres cannot be mechanically harvested this year and cannot be grazed between April 1 and September 1.

If the farmer needs forage, Hicks and his colleagues say, he would do well to establish a long-term stand of alfalfa, perhaps with a companion crop. The alfalfa could not be harvested this year for hay, but it could be grazed after Septem-



Passengers rate the Boeing 767 as a superb flying experience. It has two wide aisles and is the only wide-body with seven-abreast seating, so nearly everyone has either a window or aisle seat. Pilots rate the 21st Century flight control system as the most advanced in the world. BOEING
Getting people together.

A BERLITZ® LANGUAGE COURSE
in a handsome briefcase
for only $125
if *you* pay for it
BUT
. . . wouldn't your *company* be interested in having you speak a second language?



These days a second language can help you help your company and so help *you* earn more. The many hispanic nationals now in our work force respond more readily to someone who speaks their language. Also, selling today is international. Whatever language it would pay you to speak, Berlitz has a do-it-yourself course in a handsome briefcase you can take anywhere. Contents—90 minute introductory tape, 40 lessons on five 60-minute cassettes, 6 illustrated word and phrase books, rotary verb finder to help you find verb tenses instantly. If not satisfied, return within 10 days and get your money back. Learn from the leader!

For your convenience on credit card orders dial toll-free
1-800-228-2028 ext. 35
and **refer to Dept. 3163**
24 hours a day, 7 days a week.

Why not give us a ring—right now! (our expense)

(Nebraska residents dial 402-571-4900)

Berlitz Publications, Inc., Dept. 3163
3490 Lawson Blvd., Oceanside, NY 11572

Send **Berlitz Comprehensive Cassette Course(s)** checked. $125 each in briefcase plus $4 for shipping and insured delivery.

☐ **French** 86100 ☐ **German** 86101 ☐ **Italian** 86102 ☐ **Spanish** 86103

Enclosed check _____ money order _____ payable to Berlitz.
Or charge my AMEX _____ Diners Club _____ VISA _____
MasterCard Interbank # _____ Exp. Date _____
Card # _____ Exp. Date _____
Name _____
Address _____
City _____ State _____ Zip _____

N.Y. residents add sales tax. Allow 4 weeks for delivery.

ber 1; the companion crop would have to be made unharvestable. "We believe intensive grazing with 10 to 20 mature cows per acre between September 1 and September 15 would provide for higher-quality forage the following year," the Minnesota group writes.

If the farmer does not need forage, they say, he still should manage the diverted acres to control erosion by wind and water and to keep down weeds. "The simplest method would be to leave the 1982 corn stubble (if it is still there) and control the weeds with herbicides or by mowing." Since research at the University of Minnesota has shown that corn yields are reduced when corn is grown following a year when no cover crop was planted, the diverted acres "would be best managed by planting a legume to provide 1983 cover and improve the nitrogen status of the soil for the 1984 corn crop."

Another possibility is a grass crop. It would not add nitrogen to the soil as legumes do, but it would improve the physical condition of the soil when it was plowed under. Neither legumes nor any other crop serving as a "green manure crop" that is plowed under will add nutrients (except nitrogen). "The phosphorus, potassium and other elements in the plant growth were taken from the soil in the first place but will be available again as the green manure decomposes." Winter rye, winter wheat and winter barley are also good cover crops if they are seeded in the spring. Then they would not have to be mowed to prevent the formation of heads, that is, to be made unharvestable.

Talkwriters

The development of a machine that can automatically transcribe human speech has been painfully slow and piecemeal; the fundamental difficulty is that pronunciation varies enormously from person to person and varies widely even for a single speaker, depending on the linguistic context of a given sound, the speaker's mood, the quality of the microphone used for transmission and so on [see "Speech Recognition by Computer," by Stephen E. Levinson and Mark Y. Liberman; SCIENTIFIC AMERICAN, April, 1981]. Investigators at several major laboratories are studying the problem from a generalized perspective: To what extent is it possible to build a machine that understands language the way people do? A much less ambitious goal, however, may have wide practical application: the development of a voice-actuated typewriter. Now Raymond Kurzweil, a computer scientist formerly at the Massachusetts Institute of Technology, has formed a company that will attempt to build a voice typewriter, capable of transcribing speech drawn from a vocabulary of 10,000 words, by 1985. The company, Kurzweil Speech Systems, Inc., is independent, but it will benefit from a large investment in development by the Xerox Corporation.

Any machine for recognizing speech that is likely to be available in the next 10 years will be a compromise among cost, technical capability and the buyer's acceptance; the transcription of fast, connected speech by many speakers over a large vocabulary remains technically infeasible. Kurzweil's strategy calls for building a $5,000 machine that can respond to discrete dictation, in which utterances must be parsed by the speaker with a very brief pause between each word. Nevertheless, the machine would take dictation as fast as 150 words per minute. It would be trained to the individual characteristics of a speaker's voice for about an hour before its ordinary operation could begin. The refinement of the templates used for pattern recognition would continue, however, even after the training period, and the words that made up the ultimate 10,000-word vocabulary would be the ones most often used by the speaker. The recognition system is designed to be compatible with many existing personal computer work stations, and so the editing of a text could also be done by voice command.

Several other companies are now manufacturing voice-recognition systems for special purposes, but none so far has publicly entered the race to make a voice-actuated typewriter. Threshold Technology, Inc., which declared bankruptcy last year, is still making voice recognizers intended chiefly for noisy, industrial environments. The machines can recognize 340 words, spoken with a short pause between each word, at up to 180 words per minute. The company expects soon to announce a device that can recognize up to 1,500 words spoken in isolation, and the machine will recognize connected utterances of any length drawn from a preselected vocabulary of about 50 words.

One of the most sophisticated devices commercially available for recognizing English is made in Japan by the NEC Corporation. The machine has to be trained by the individual speaker, but it can recognize connected utterances up to four seconds long drawn from a 150-word vocabulary. Other machines are being developed in Japan that are specially adapted to the Japanese language, and their capabilities cannot be readily compared with those of machines designed for recognizing English. There are only 120 syllables in Japanese, compared with some 10,000 in English, and the Japanese speaker is much more regular in pronunciation, with fewer glides from syllable to syllable, than the English speaker. Moreover, the incentive for the development of a voice-actuated typewriter is probably stronger in Japan than it is in the U.S. and Europe because the potential benefit is much greater: the Japanese typist must select from more than 3,200 characters.

Workers at Bell Laboratories have almost completed the development of several speech-recognition products that will be announced at the end of the year. Bell is known to be designing systems that can respond to many speakers as well as systems that must be individually trained. Speaker-independent systems would enable any person to control a computerized data bank, place an order or respond to a programmed series of questions over the telephone without the intervention of another person. The tradeoff is that the recognized vocabulary must be smaller than it is on a machine designed to recognize one speaker at a time.

At the International Business Machines Corporation the emphasis has been on developing an office dictation system that can handle connected speech drawn from a vocabulary of at least 5,000 words. According to Frederick Jelinek, who heads the speech-recognition project at IBM's Thomas J. Watson Research Center, the aim is to develop a machine that can be trained to the voice pattern of the individual in less than 15 minutes. The approach to pattern matching is a statistical one: the spectrum of the speech event is sampled every 10 milliseconds and each sample is assigned to the most similar of 200 spectral patterns digitally stored in the memory of a computer. Words are recognized by finding the word that most closely matches a particular string of spectral patterns and also has a high probability of following the two preceding words; the probabilities related to three-word sequences are derived from extensive analysis of business correspondence. According to Jelinek, IBM is not yet developing any speech-recognition machine for commercial use, but the company is clearly in the race to perfect a voice-actuated typewriter.

There is good reason, in spite of the commitment of other companies, to take Kurzweil's entry seriously. In 1975 he built the first working model of a machine that can now recognize printed words in more than 200 type fonts and then read them aloud. The most important use for the device so far has been as a reading machine for the blind. A blind person can place a book, newspaper or any other printed text on top of a glass window and then listen as the machine automatically reads the text in a synthesized voice. The voice can be speeded up to 225 words per minute, and the user can ask the machine to repeat a phrase, spell out a word or skip sections of text. Kurzweil Computer Products, Inc., a subsidiary of Xerox, now markets the reading machine primarily to libraries and other institutions for about $30,000.

Although optical character recognition and the recognition of vocal sounds make quite different technical demands, the underlying problems are the same: an enormous number of patterns must be compared with a limited number of internal models in a short time.

Chemical Pruning

The fruit produced by an orchard tree represents an investment of energy by the tree, and the less energy the tree has to put into new vegetative growth, including leaves, the more energy it can put into fruit and the better fruit it produces. Moreover, when an orchard tree has too many leaves, sunlight cannot reach some of the places where flowers (the first stage of fruiting) might appear and cannot reach the fruit effectively enough to bring the fruit to its full ripe color. Orchardists currently deal with the excess-foliage problem by pruning, which is both labor-intensive and costly. Now there is a chemical to do the job. It should be available commercially within a few years as a result of work by plant physiologists at the Fruit Research Laboratory of the Agricultural Research Service of the United States Department of Agriculture in Wenatchee, Wash.

The chemical has the proposed common name of Paclobutrazol but at present is designated ICI pp333. It was developed to control the growth of grass so that farmers whose product is grass seed would get a higher yield. Applied to the ground around a fruit tree, the chemical holds down the growth of new leaves on the tree. The result is a tree that looks as though it had been skillfully pruned by hand. Because of the relative lack of foliage, the tree has a large crop of fruit of uniform size and good color. The work is characterized by Max W. Williams, director of the project: "We are using chemicals to accomplish what genetics has so far failed to do, and for the first time ever we can have control of an apple or pear tree's vegetative and fruit growth throughout the life of the tree."

Williams and his colleagues began testing the compound on fruit trees as a foliage spray in 1978. In 1979 they started applying it to the ground at the rate of 2.5, 5, 10 and 20 grams of active chemical ingredients per tree. Williams foresees that in commercial practice the rate of application might be from one gram to five grams per tree. The ground treatment is superior to the foliage spray and avoids chemical contamination of the fruit. It has worked successfully with several kinds of apple and with Anjou pears, and Williams believes it will succeed with peaches and cherries. The chemical treatment may also help in the control of insects and diseases that prey on the trees, since leaves can sustain a variety of insects and fungi.



World travelers choose the 737 because it has one of the best on-time records. The 737-300 is the latest addition to the family. This new jetliner, with added passenger amenities, advanced flight controls and quieter engines, will make flying an even better travel experience.

BOEING
Getting people together.

In 1982, Porsche won the World Championship of Makes. Audi won the World Rally Championship for Manufacturers and SCCA Pro Rally Championship.

What we learn from our competition cars, we put into our production cars.

And now you can arrange to drive your new Porsche or new Audi where it was built, tested, and refined—with our Delivery In Europe Program.

Simply visit your local Porsche Audi dealer and order the model of your choice: the Porsche 928S, 911SC, or 944. Or the new Audi 5000S, the 4000S, Coupe GT, or all-wheel drive Quattro. Then take delivery in Europe. And after your stay, we'll deliver your car to a U.S. port free of charge.

For more information, see your Porsche Audi dealer: the home of champions.

Visit Europe And Drive Home A World Champion.



The world's densest computer is now the heart of the world's most 32-bit computer.

Our 32-bit CPU.

From time to time, miracles of technology come along to make previously impossible tasks not only possible, but easy. That tiny 450,000-transistor integrated circuit is one of those technological miracles.

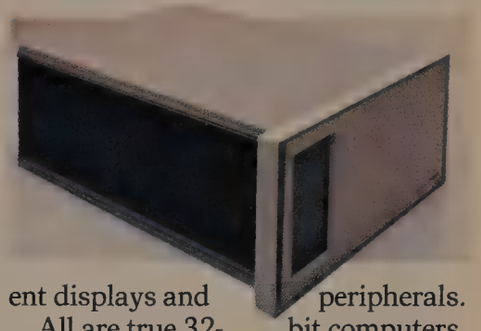
Hewlett-Packard didn't develop it just to break the record for most transistors on a chip, but to put on an engineer's or scientist's desk a computer so powerful that it can do the work of mainframes costing four times as much.

32-bit computers for 32-bit applications.

The new HP 9000 computer based on this and four other 'superchips' can handle formidable engineering and scientific problems. The scientist solving complex systems of equations, the mechanical engineer doing finite element analysis or three-dimensional modeling, the electrical engineer analyzing complex circuits or designing very large-scale integrated circuits—these are the kinds of technical people and problems the HP 9000 family is designed for.

It comes in three versions. The integrated workstation is complete with keyboard, color or monochromatic graphics display, fixed and flexible disc drives, and printer. For systems manufacturers, there's a rack-mountable box. And for a variety of single-user and multi-user applications, the minicabinet version works with many different displays and peripherals.

A rack-mountable version is available, too.



All are true 32-bit computers, with 32-bit CPUs, memories, and data paths. And the multi-CPU architecture lets you nearly double or triple your processing power at any time by adding one or two CPU boards. Without increasing the computer's size.

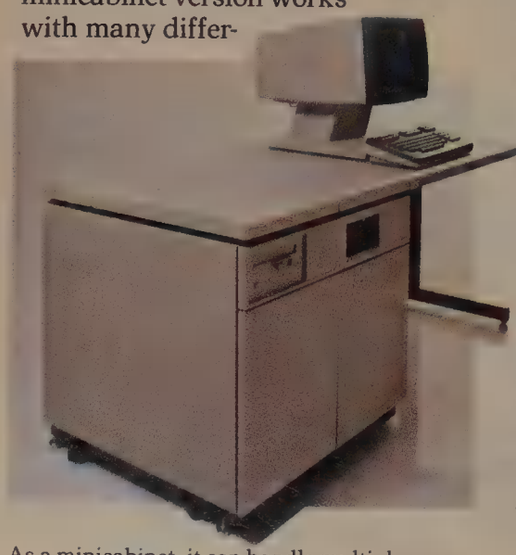
Two operating systems are better than one.

The integrated workstation is available with a choice of operating systems. One is HP's highly evolved, high-performance Enhanced BASIC, augmented with 3-D graphics and a software innovation called a run-time compiler. This substantially increases program execution speed, while retaining an interactive development environment.

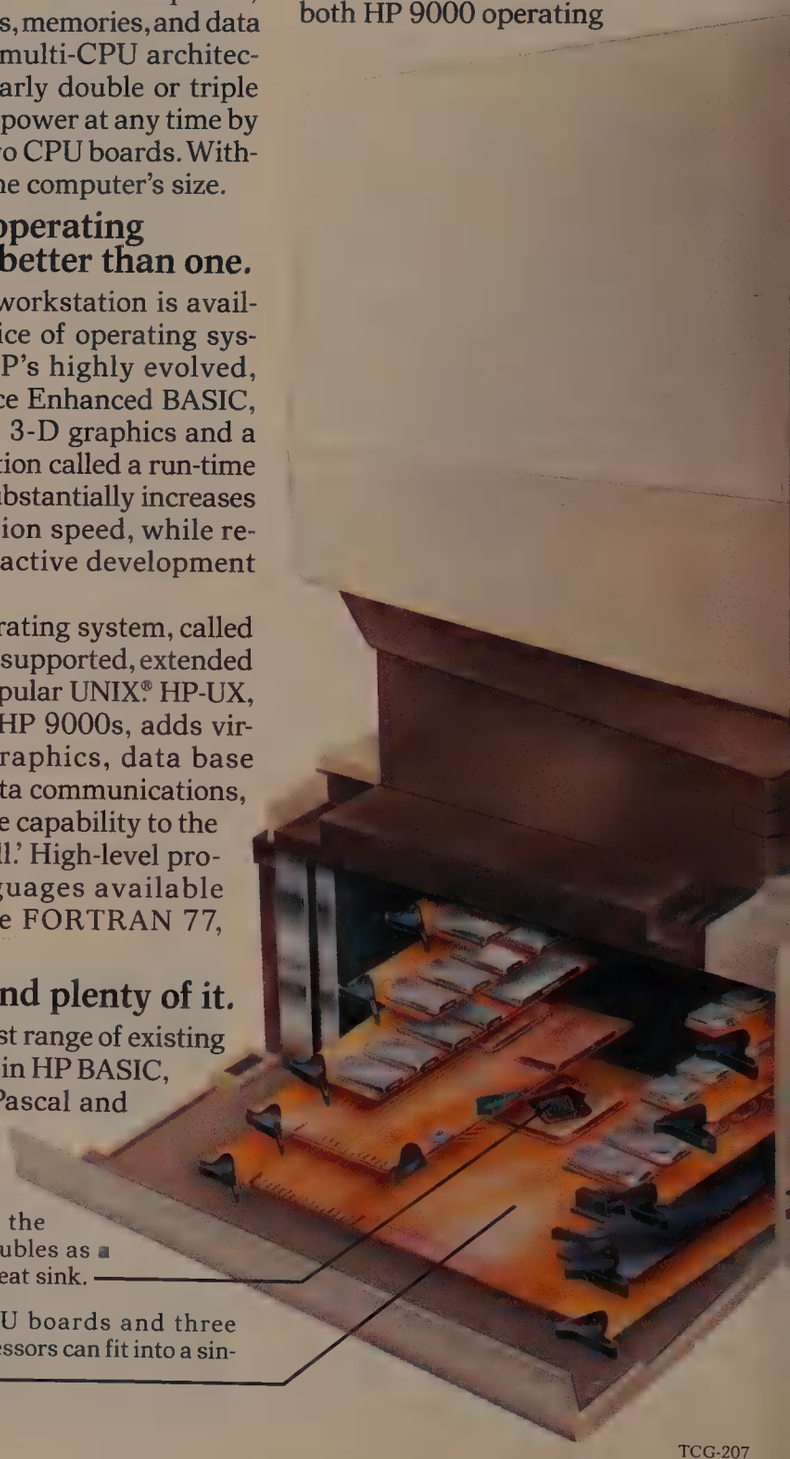
The other operating system, called HP-UX, is a fully supported, extended version of the popular UNIX® HP-UX, available on all HP 9000s, adds virtual memory, graphics, data base management, data communications, and enhanced file capability to the basic UNIX 'shell.' High-level programming languages available with HP-UX are FORTRAN 77, Pascal and C.

Software, and plenty of it.

Much of the vast range of existing software written in HP BASIC, FORTRAN 77, Pascal and C is transportable to the HP 9000. HP will also be offering proprietary software packages emphasizing computer-aided design and engineering. These will tie the HP 9000 into HP's Manufacturer's Productivity Network (MPN). Third-party software suppliers will be providing many of the most widely used CAE packages for 32-bit computer systems. And both HP 9000 operating



As a minicabinet, it can handle multiple users.



The 32-bit CPU chip is bonded to the finstrate which doubles as a signal carrier and heat sink.

Up to three CPU boards and three Input/Output Processors can fit into a single HP 9000.

TCG-207

ip
fordable

systems offer extensive program development tools.

You also get a choice of communication tools. The HP 9000 is currently compatible with Ethernet,™ and with HP's Shared Resource Manager (SRM) which lets clusters of HP 9000 and 16-bit desktop computers share data and use common peripherals.

Links to central computers are also available. And in late 1983, HP will offer local area networks based on the IEEE-802 standard.

New technology from the silicon up.

The five superchips that make the HP 9000 possible are the 32-bit CPU, which can execute a million instructions per second; an eight-channel Input/Output processor (IOP); a random-access memory chip capable of storing 128K bits of data; a memory controller that 'heals' up to 32 bad memory locations; and an 18-megahertz clock.

Hewlett-Packard's advanced NMOS-III process makes it possible to put 450,000 transistors on a chip only 0.4 centimeters square. This tremendous density of electronic components could have required an expensive and elaborate cooling system.

Instead, HP engineers developed a new mounting structure called a finstrate, a copper-cored circuit board, which acts as both cooling fin and substrate. The finstrates containing the CPU, IOP, memory, and clock chips are housed in a lunchpail-sized module.

One user, one mainframe.

Clearly the trend in engineering and scientific computation is away from large machines shared by multiple users and towards networks of powerful personal workstations, sharing peripherals and data bases. The reason is compelling. An engineer or scientist in personal control of an HP 9000 can solve so many more problems more easily that the increased productivity alone makes the cost of individual computers easy to justify.

For complete information about this powerful breakthrough in 32-bit computing, contact the local HP sales office listed in your telephone directory. Ask a Technical Computer Specialist for a demonstration. Or write to Pete Hamilton, Dept. 41151, Hewlett-Packard, 3404 East Harmony Road, Fort Collins, CO 80525.



HEWLETT PACKARD

UNIX is a trademark of Bell Laboratories. Ethernet is a trademark of Xerox Corporation.

Interstellar Matter in Meteorites

Carbonaceous chondrites, the most primitive meteorites, incorporate material originating outside the solar system, including matter expelled by supernovas and other stars

by Roy S. Lewis and Edward Anders

The solar system is much younger than the universe (only 4.5 billion years compared with 10 to 15 billion years), and so it must have formed from older matter that had a previous history. Before 1969, however, no such relict matter had ever been found, either in meteorites or in planets. And since theorists had proposed a hot origin for the solar system, which would have caused all preexisting solids to vaporize, it seemed that the solar system started out with a clean slate.

How could one recognize presolar matter? Perhaps by its age, but more reliably by its isotopic composition. Stars continuously rework the chemical elements and then eject them back into interstellar space, where they eventually form the next generation of stars. Both the elemental and the isotopic composition of the ejected matter vary from star to star, depending on the star's mass, temperature and stage of evolution. The isotopic composition is a particularly durable hallmark, since it can be changed by few processes (short of nuclear reactions) and then only to a limited, predictable extent. Therefore one looks for matter of unusual isotopic composition that cannot be explained by known or plausible processes within the solar system.

The first hint of presolar matter came, then, in 1969, when David C. Black and Robert O. Pepin of the University of Minnesota studied the isotopic composition of the noble (chemically inert) gas neon in carbonaceous chondrites: dark gray, rather nondescript-looking stones that are the most primitive meteorites known. Hidden by large amounts of normal neon was a small component of neon greatly enriched in the isotope neon 22. Given the rarity and oddity of noble gases, a local origin could not be ruled out. Then in 1973 Robert N. Clayton, Lawrence Grossman and Toshiko Mayeda of the University of Chicago found that certain minerals in the Allende carbonaceous chondrite were enriched in the isotope oxygen 16 by as much as 5 percent with respect to normal oxygen. After ruling out a local origin they concluded that some oxygen made by nuclear reactions in stars other than the sun had found its way into the meteorite.

Other discoveries followed, and today some 19 chemical elements in carbonaceous chondrites are known to show isotopic anomalies. Evidently the slate was not quite clean; some presolar matter did survive in the cooler precincts of the early solar system. The challenge is to find these bits of presolar matter and to decipher their record of stellar nucleosynthesis and interstellar chemistry.

Carbon has yielded a particularly rich crop of presolar components; at least four have been recognized so far. All are well hidden and were found only because they are "tagged" with anomalous noble gases or hydrogen. Let us follow the four trails of discovery, after first describing the meteorites in which the anomalies are found.

Carbonaceous Chondrites

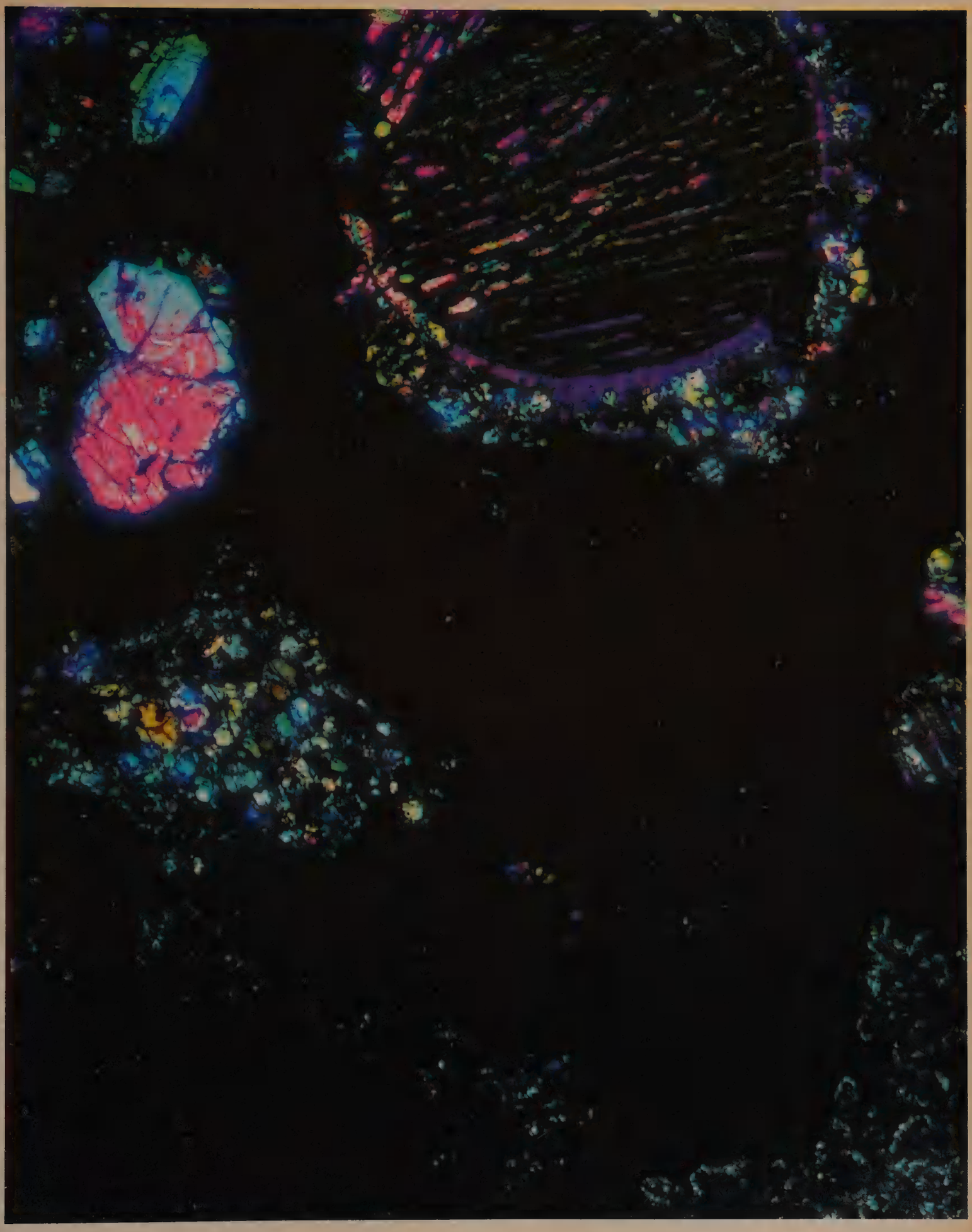
About 40 carbonaceous chondrites are known; they represent some 4 percent of all the meteorites observed to fall. They are divided into three main classes, designated C1, C2 and C3, on the basis of mineralogy and declining content of volatiles such as water, carbon and nitrogen. Their formation temperature (that is, the temperature at which they ceased to react with the gas in the solar nebula, the cloud from which the solar system condensed) increases from C1 to C3, so that if the temperature increased toward the center of the nebula, the C1's come from farthest out. (They formed at a temperature of 360 degrees Kelvin. C3's formed at a temperature some 60 degrees higher.) All carbonaceous chondrites were altered in their parent bodies, probably asteroids. C1's and C2's were exposed there to liquid water, and C3's were somehow reheated to a temperature of perhaps 600 degrees K.

Texturally the carbonaceous chondrites consist of a fine-grained matrix in which coarse-grained particles are embedded. The matrix particles range in size from 10^{-4} to 10^{-6} centimeter and are rich in volatiles; they consist of silicates and carbonaceous matter. The larger particles range in size from .01 centimeter to one centimeter. They are poor in volatiles. Some of them are chondrules: rounded particles looking much like bird shot. Others are more irregular but consist of the same minerals: olivine [$(Mg,Fe)_2SiO_4$], pyroxene [$(Mg,Fe)SiO_3$], troilite (FeS) and nickel-iron. Still others are refractory (heat-resistant) inclusions of minerals rich in calcium, aluminum and titanium but poor in silicon. In retrospect it is not surprising that isotopic anomalies are found in the matrix, which has been heated very little, and in the refractory inclusions, which are highly resistant to heat. The anomalies we shall now discuss are found chiefly in the matrix.

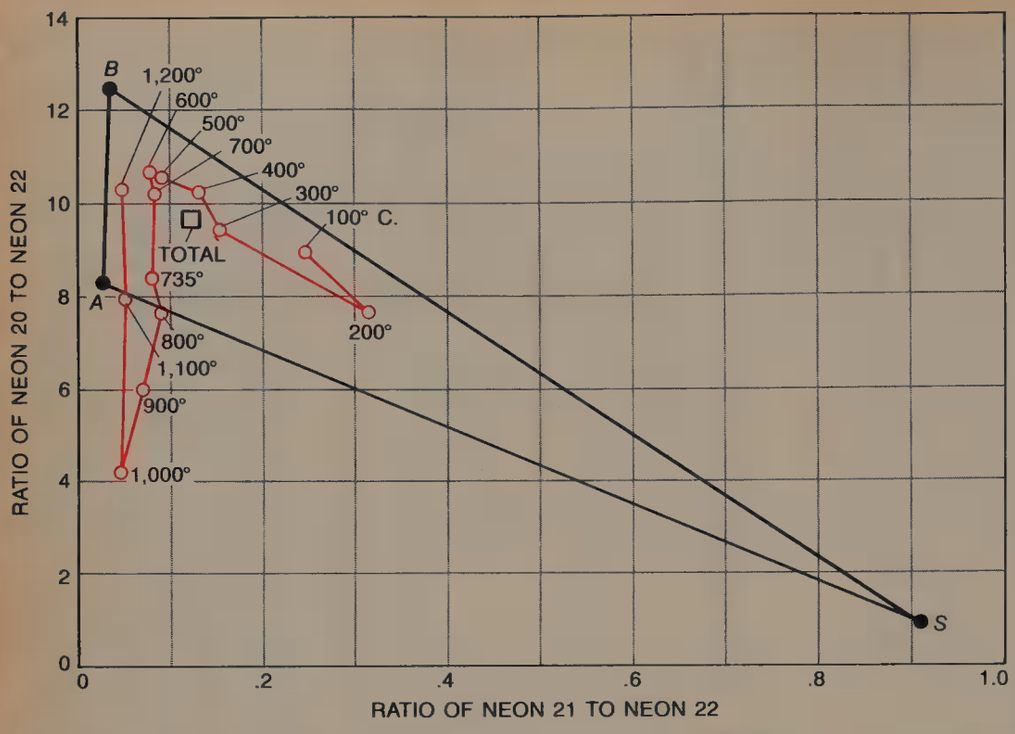
Neon *E*

We shall begin with the noble gases. Among the elements in meteorites they are unique. Being highly volatile and unreactive, they did not condense in even the most primitive meteorites and hence are present at only a minute fraction of their proportion in the sun, ranging from about 10^{-5} for xenon to 10^{-9} for helium and neon. This tiny amount of gas, however, is tightly bound in the meteorite, coming free only at high temperatures when its host mineral begins to melt or decompose.

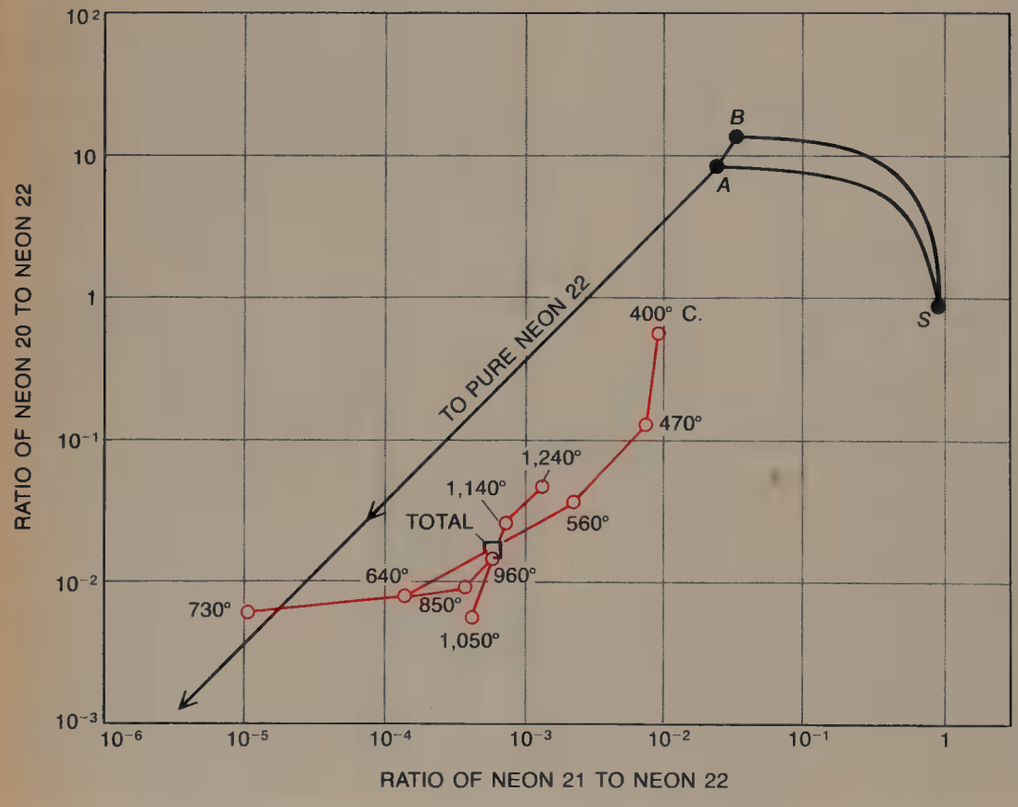
Black and Pepin tried to untangle three types of neon known to be present in primitive meteorites: primordial or planetary neon, also called neon *A*, which was trapped from the solar nebula; solar neon, also called neon *B*, which consists of solar-wind neon ions implanted in meteorites that happen to have been at the surface of their parent body; and cosmogenic neon, also called neon *S*, formed when cosmic rays pass-



THIN SECTION of the Allende meteorite photographed under polarized light shows the characteristic texture of carbonaceous chondrites. The field of view is about a millimeter across. At the upper right is a chondrule, a solidified silicate droplet. Elsewhere in the field are irregular bits of silicates, sulfides and nickel-iron. At the lower right is a refractory (high melting point) inclusion of the mineral olivine. The rest of the field, black in the photograph, is a fine-grained matrix consisting of silicates and carbonaceous matter. Proportions of isotopes unlike those of the solar system's matter have now been found for some 19 chemical elements in carbonaceous chondrites, mostly in inclusions and in the matrix. The photograph was made by Glenn J. MacPherson of the University of Chicago.



THREE-ISOTOPE CHART reveals anomalous neon in the Orgueil meteorite. A sample of the meteorite has been heated to progressively higher temperatures, and the fraction of gas released at each temperature has been analyzed to determine its proportions of the isotopes neon 20, 21 and 22. The first two fractions are well within the triangular area representing mixtures of the three types of neon expected in meteorites: primordial neon (*A*), which comes from the solar nebula, the gas cloud that formed the solar system; solar neon (*B*), implanted in the meteorite by the wind of ions flowing outward from the sun, and cosmogenic neon (*S*), manufactured by collisions between cosmic rays and atomic nuclei inside the meteorite. Then, however, a sequence of fractions veers out of the triangle. Evidently something in the Orgueil meteorite releases a type of neon greatly enriched in neon 22. It was later called neon *E*. The experiment was done in 1969 by David C. Black and Robert O. Pepin of the University of Minnesota.



IN A LATER EXPERIMENT M. H. A. Jungck and Peter Eberhardt of the University of Bern applied stepped heating and three-isotope charting to a tiny mineral sample greatly enriched in neon *E* that they prepared from the Orgueil meteorite. The neon *E* proved to be essentially pure neon 22 (more than 99 percent for the fractions released from 640 through 1,050 degrees Celsius). Apparently neon *E* formed from the radioactive decay of sodium 22 in grains around a star, perhaps a nova. When the solar system condensed, the grains were incorporated into the meteorite. The chart is logarithmic; the triangle defined by *A*, *B* and *S* is at upper right.

ing through the meteorite spall, or shatter, atomic nuclei in their path. Each type has different proportions of the three isotopes of neon.

Black and Pepin employed the technique of stepped heating. Here a sample of the meteorite is heated to progressively higher temperatures, and the gas released is analyzed in a mass spectrometer. With luck the different gas components emerge one by one as their host minerals melt, decompose or become permeable. More often the gases emerge as a mixture that must be further resolved on the basis of isotopic composition. This can be done most easily by plotting two isotopic ratios against each other, both with the same denominator. In the case of neon one plots the ratio of neon 20 to neon 22 against the ratio of neon 21 to neon 22 for each temperature step. On such a "three-isotope plot" all mixtures of two types of neon lie on a straight line joining the two, mixtures of three types lie within a triangle whose corners are the three, and so on.

All meteoritic neon samples measured up to the time of Black and Pepin's work had fallen within the triangle bounded by neon *A*, neon *B* and neon *S*, as expected for mixtures of those three. When Black and Pepin analyzed six C1 and C2 chondrites by stepped heating, they found that the fractions released between 800 and 1,100 degrees Celsius consistently fell below the triangle. Evidently a new neon component was present, with a ratio of neon 20 to neon 22 that was less than 3.4. They named it neon *E*, the letters *C* and *D* having been preempted by two minor components of less extreme composition. Black and Pepin pointed out that the neon *E* could not have derived from solar neon by processes of mass fractionation such as diffusion through a solid or gravitational escape from a planet. The initial reservoir of neon would have been absurdly great: some 10^{10} times greater than the estimated content of neon in the solar nebula.

An alternative possibility is the intense irradiation of dust grains by protons (hydrogen nuclei) in the early solar system. Spallation reactions between protons and magnesium in the grains could make not only the isotopes of neon in the proportions of neon *S* but also sodium 22. If the grains were heated during the irradiation or right after it, the neon *S* would escape. The sodium 22 would then decay, with a half-life of 2.6 years, to neon 22. Donald D. Clayton and his co-workers at Rice University have noted, however, that proton irradiation should also have produced excesses of argon 36 and krypton 80. Such excesses are not observed.

That leaves stellar nucleosynthesis. Stars that have exhausted their hydrogen build neon and neighboring elements by thermonuclear reactions in-

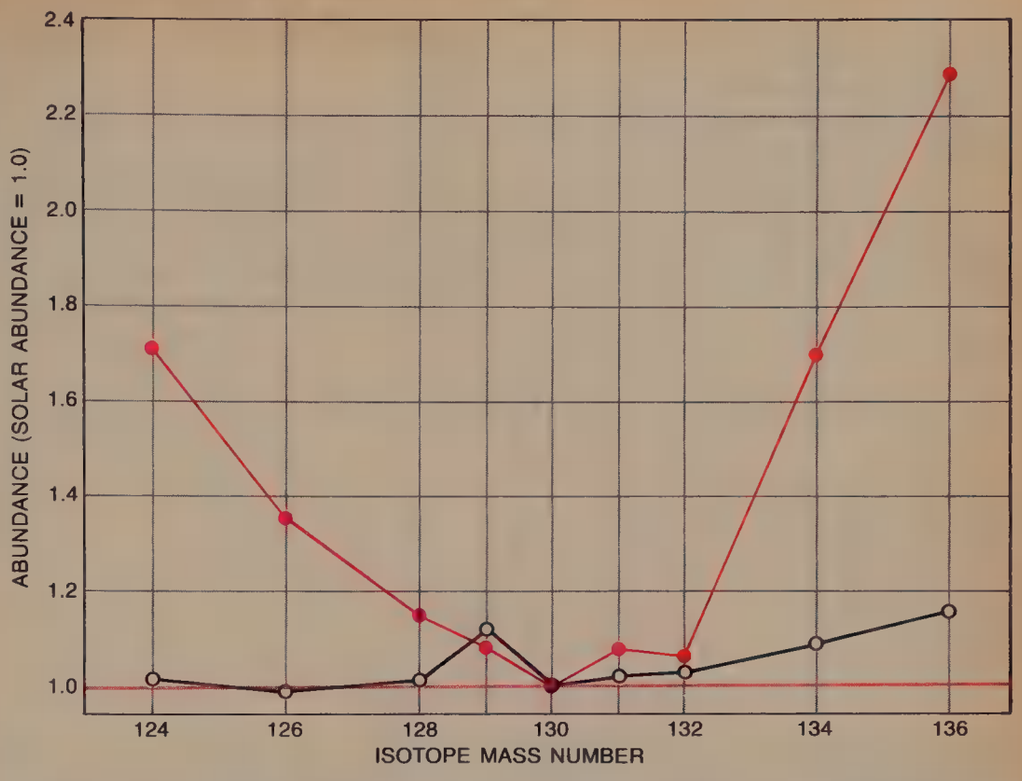
volving helium. For example, helium 4 combines with oxygen 16 to make neon 20. M. Arnould of the Free University of Brussels and H. Nørgaard of the Nordic Institute for Theoretical Atomic Physics (NORDITA) in Copenhagen have shown that neon *E* could form in stars, either directly (under special conditions that greatly favor neon 22 over the other isotopes of neon) or indirectly (as sodium 22 that preferentially condenses on dust grains after ejection from the star and then decays into neon 22).

The Carriers of Neon *E*

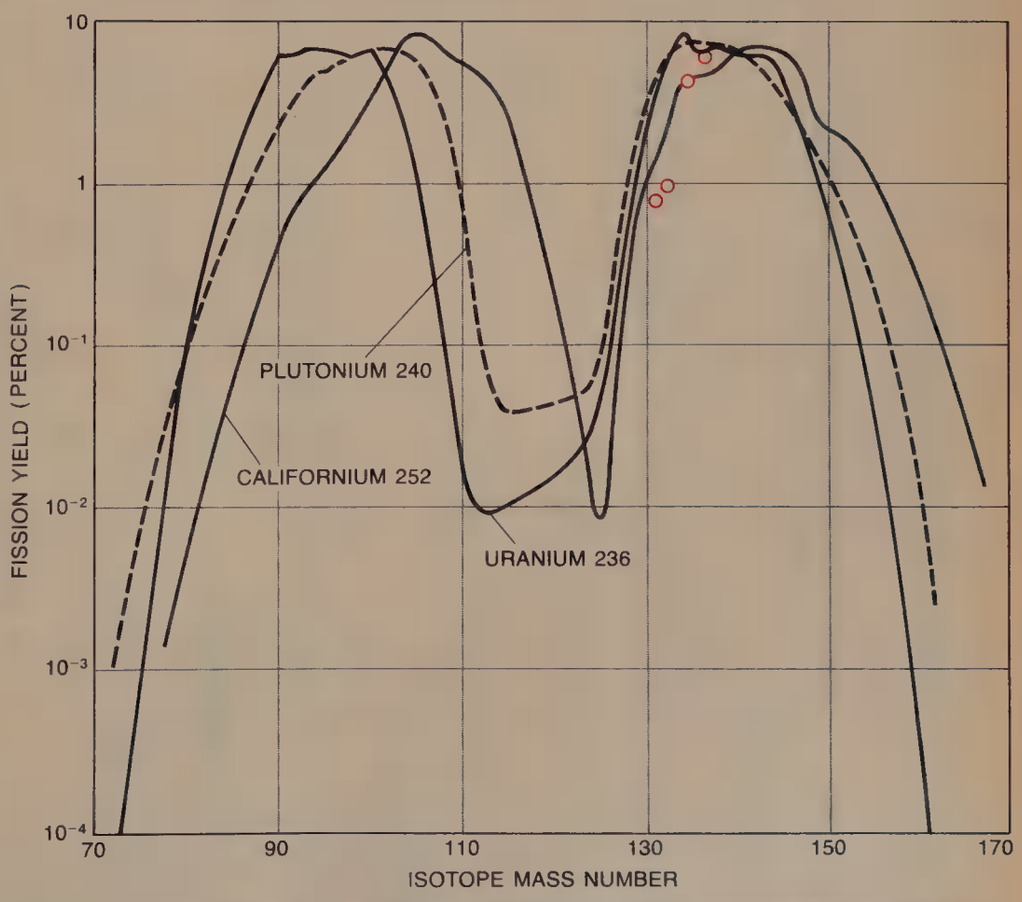
Here was a chance to study "stardust," and perhaps even to learn what type of star it came from. The obvious next step was to isolate the host mineral (the "carrier") of the neon *E* and learn whether very pure neon *E* contained only neon 22 or the other two isotopes of neon as well. This problem was solved mainly by Peter Eberhardt of the University of Bern and his students F. Niederer, M. H. A. Jungck and F. O. Meier. Starting at Chicago in 1974, Eberhardt systematically took the Orgueil meteorite apart by various physical methods of separation. The mineral magnetite, for example, could be removed with a magnet. Fine-grained silicates could be removed by letting them form a colloidal suspension in concentrated salt solutions. At each stage Eberhardt checked to see where the neon *E* had gone. In this way he found that it concentrated in a coarse-grained, nonmagnetic fraction. Further separations showed that the neon *E* had at least two carriers, one of low density and release temperature (less than 2.2 grams per cubic centimeter and 600 degrees C.), the other of higher density and release temperature (3.5 grams per cubic centimeter and 1,100 degrees). The corresponding types of neon *E* were named *L* for low and *H* for high.

Further experiments at Bern and Chicago (in which Leo B. Alaerts took part) established the nature of the carriers. The carrier of neon *E*(*L*) is carbonaceous. The carrier of neon *E*(*H*) consists of two minerals: spinel ($MgAl_2O_4$) and apatite [$Ca_2PO_4(OH,F)$]. Meanwhile the compositional limits for neon *E* were being tightened from year to year, strengthening the possibility that neon *E* is pure neon 22 and hence derived from sodium 22. The issue was clinched by Jungck and Eberhardt, whose purest sample was nearly 99 percent neon 22.

Now that we know the carriers of neon *E*, what can be said about its origin? The short, 2.6-year half-life of sodium 22 means it formed quickly and was trapped quickly, before it decayed to neon. Donald Clayton has pointed out that these conditions are met by exploding stars: novas and supernovas. Both synthesize elements explosively and



ANOMALOUS XENON was discovered in the Renazzo meteorite by John H. Reynolds and Grenville Turner of the University of California at Berkeley; in a later effort its isotopic composition in the Allende meteorite was analyzed by B. Srinivasan and the authors at Chicago. First a sample of carbon grains from Allende was analyzed; its xenon (*black*) differed only slightly in composition from solar xenon. Then the investigators etched the surface of the grains with nitric acid. The etching removed some 95 percent of the xenon. The remainder (*color*) was grossly enriched in both heavy and light isotopes. Evidently the xenon in the sample was a mixture of primordial xenon on the surface of the grains and two types of anomalous xenon inside. One type (xenon *H*) is enriched in heavy isotopes, the other (xenon *L*) in light isotopes.



FISSION OF HEAVY ELEMENT inside meteorites was the first proposal for the origin of xenon *H*. It was known that the nuclei of uranium and other heavy radioactive metals split asymmetrically, yielding the two-hump distributions of fission products shown in the illustration. The heavy isotopes of xenon discovered by Reynolds and Turner (*colored dots*) fit this general trend but do not match any one known curve in detail. Reynolds and Turner therefore suggested that xenon *H* might come from the fission of some unknown superheavy element now extinct in the solar system. Xenon *L*, however, would have formed some other way.

eject them almost instantaneously. Interestingly, the first minerals predicted to condense from the gases ejected by these two types of stars are respectively carbon and spinel, the carriers of neon $E(L)$ and neon $E(H)$. Only one part in 10^9 of the mass of each grain would have to be sodium 22 in order to account for the neon E today. The carrier apatite, however, remains unexplained. It is not an expected condensate in a stellar explosion, or in the solar nebula for that matter.

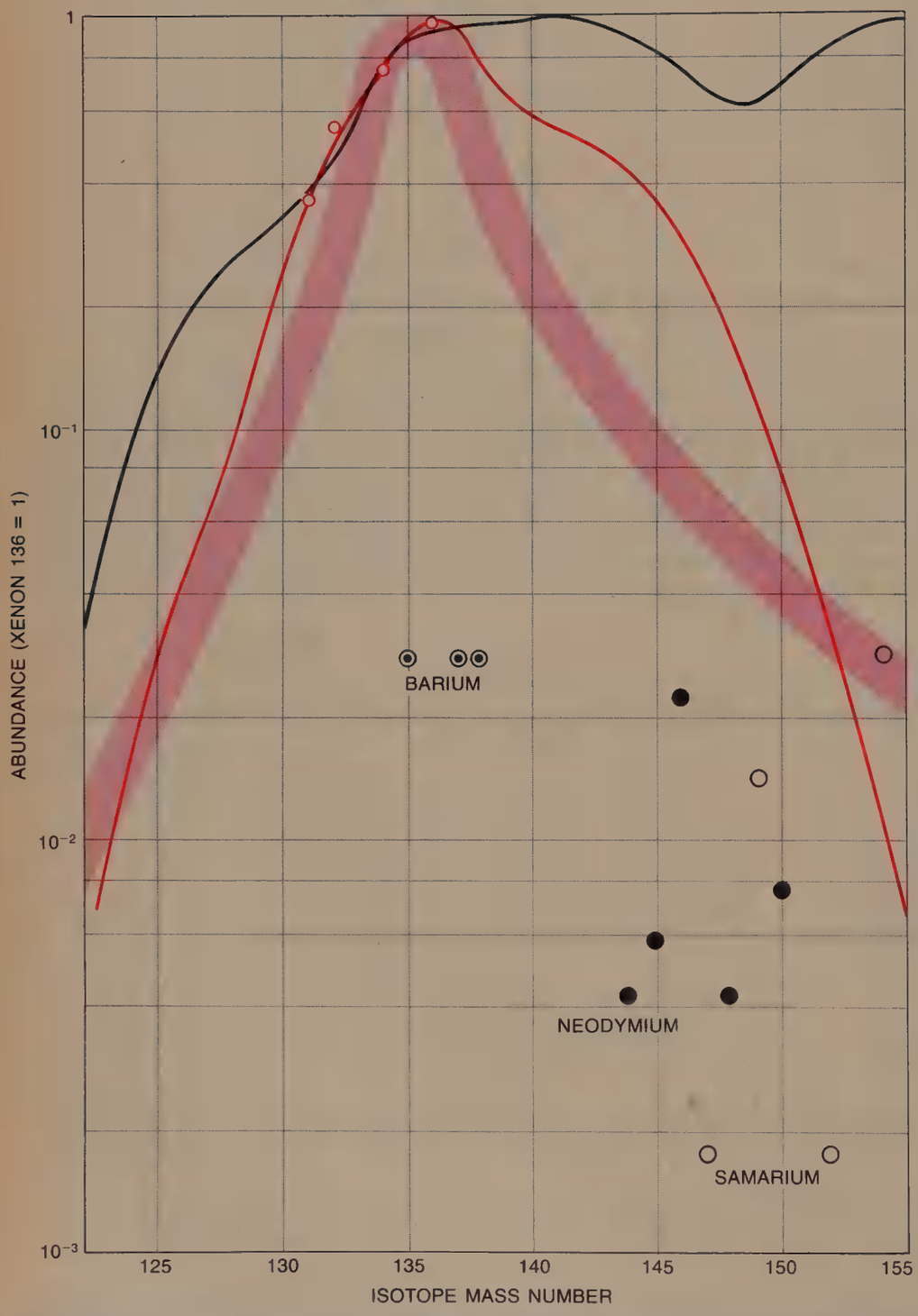
Xenon *HL*

When William Ramsay, having discovered xenon in 1898, named it after the Greek word for stranger, he could not have foreseen how well the name would fit the xenon in meteorites. At least three strange types of xenon are present in carbonaceous chondrites. Two are abundant but controversial, the third is rare but straightforward.

The story begins in 1964, when John H. Reynolds and Grenville Turner of the University of California at Berkeley examined the C2 chondrite Renazzo. They were looking for xenon 129 from the radioactive decay of iodine 129, an isotope discovered by Reynolds whose half-life of 16 million years means that it is extinct in the solar system. (In the 4.5-billion-year history of the solar system it has decayed through about 280 half-lives and has therefore been depleted by a factor of 10^{85}.) To get rid of the large amounts of normal planetary xenon they employed the technique of stepped heating.

They indeed found the xenon 129. In addition, however, they noticed that in the fractions released between 600 and 1,100 degrees C. the heavy isotopes of xenon, ranging in mass from 131 through 136, were enriched by as much as 6 percent with respect to primordial xenon.

The enrichment increased from isotope 131 to isotope 136, as it does in the xenon formed by the fission of uranium and other heavy elements. Reynolds and Turner therefore suggested that the new xenon component came from the fission of some extinct heavy element that had once been present in the meteorite. They also noted a problem: the rare light isotopes xenon 124 and 126 were also enriched in the meteorite. These isotopes do not form by fission, and so their enrichment "would have to be explained by some additional process." Apparently not one but two new xenon components are present in the meteorite. They were later named *H* and *L* for heavy and light. Although of different origins, they have proved inseparable in meteorites, and we shall therefore refer to their mixture as xenon *HL*.



COMPETING HYPOTHESES for the origin of xenon *H* are diagrammed. The hypothesis that it formed by the fission of a superheavy element inside the meteorite is represented by a fission-yield curve calculated for element 114 (*black*). A competing hypothesis is that the xenon was made in a supernova, either directly, when lighter nuclei captured neutrons (*light colored band*), or indirectly, when short-lived superheavy nuclei in the supernova decayed by fission (*solid color*). Both hypotheses account more or less well for the isotopes in xenon *H* (*colored dots*). In both hypotheses, however, the elements barium, neodymium and samarium should be made in similar quantity. Actually they are not; their upper limits (*black dots*) in an Allende residue rich in xenon *H* fall far below the theoretical curves. The supernova hypothesis escapes this predicament by assuming that the barium, neodymium and samarium expelled from a supernova condense on grains at a time when xenon, which is more volatile, cannot yet be trapped.

A Superheavy Element?

To be sure, none of the known fissionable elements heavier than uranium yields xenon with the isotopic pattern of xenon *H*. Moreover, the extension of the periodic table to elements beyond 100 showed that their half-lives get shorter with increasing atomic number (that is, the number of protons), and this suggested that the synthesis of new chemical elements might not be feasible beyond element 106 or 107. In 1966, however, W. D. Myers and W. J. Swiatecki at Berkeley tried to calculate the properties of still heavier elements. They came up with a surprising result: the predicted half-lives grew longer rather than shorter. According to their calculations, the next "magic numbers" (which correspond to particularly stable configurations of neutrons or protons in a nucleus) would come at 114 protons and 184

neutrons, giving rise to an "island of stability" in the periodic table centered on element 114. Efforts to synthesize such "superheavy" elements in accelerators or to identify them in cosmic rays are continuing. Their predicted half-lives, although highly uncertain, range up to 10^8 years. Thus in 1969 three groups of workers in the U.S. and Poland proposed that an extinct superheavy element in the island of stability might be the parent of xenon *H*.

Soon the idea was challenged. O. K. Manuel and his colleagues at the University of Missouri examined xenon from three carbonaceous chondrites. In every one they found the puzzling correlation of light and heavy isotopes that Reynolds and Turner had noted. That was a major embarrassment for the fission hypothesis. If xenon *H* was made by the fission of a heavy element inside the meteorite, why was it always accompanied by the same proportion of xenon *L*, which must have a different origin?

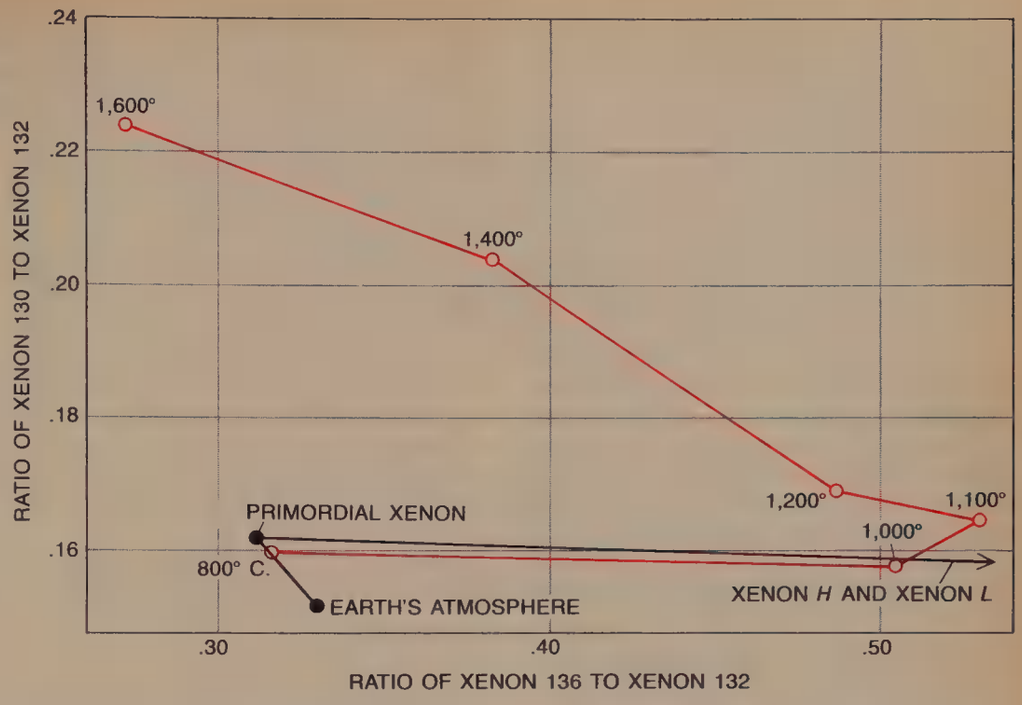
Manuel and his colleagues concluded that the two types of xenon were made together before meteorites ever formed and then were trapped in meteorites premixed. They suggested three possible origins, one of which was nucleosynthesis in different zones of a supernova. The xenon *H* would be made by the *r* process (the capture of neutrons by nuclei on a rapid time scale) and the xenon *L* by the *p* process (proton capture or some other process that gives rise to isotopes poor in neutrons).

Stumbling on the Carrier

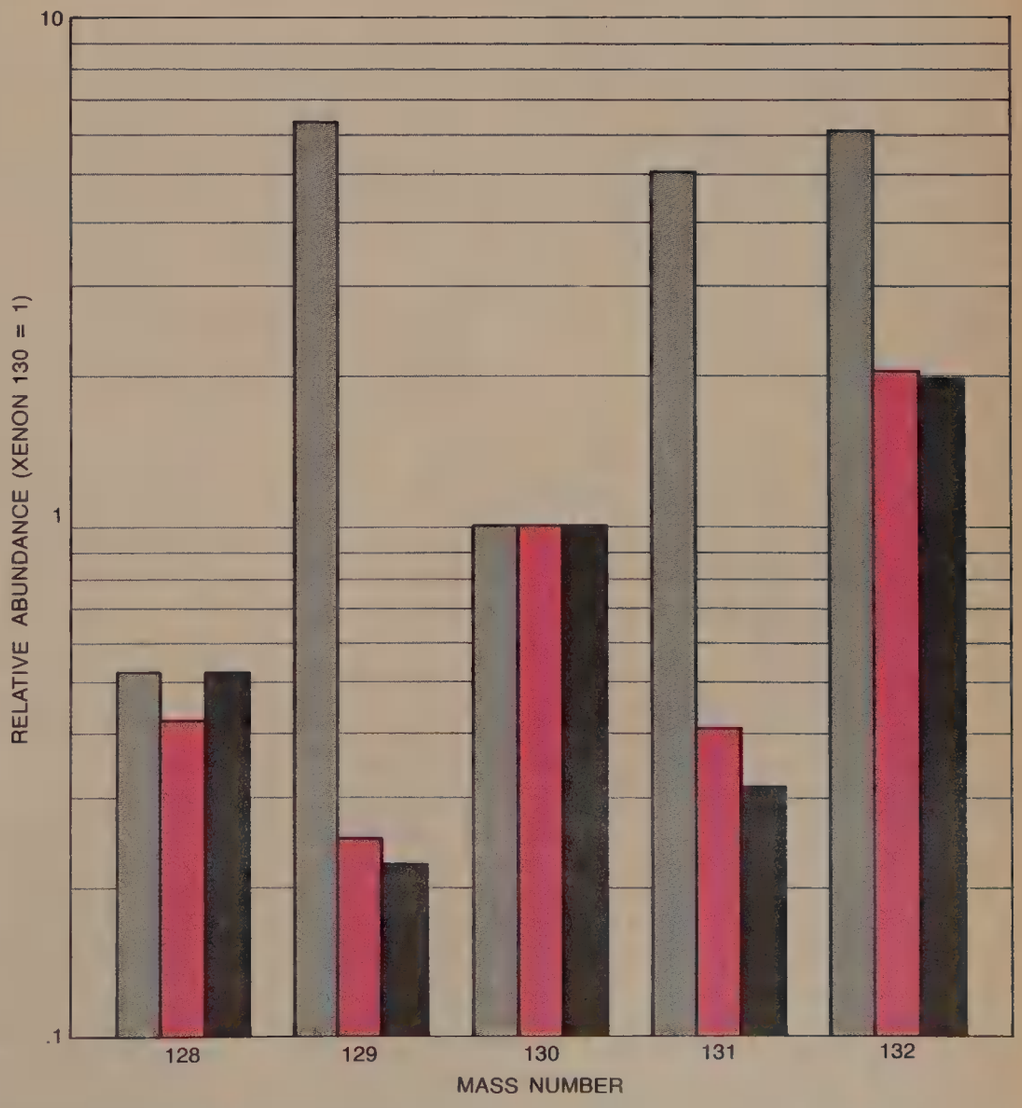
To test the hypotheses (fission inside the meteorite on the one hand, stellar nucleosynthesis on the other) it was necessary to isolate the carrier and study it in detail. If xenon *H* is exotic, its carrier should also be exotic and should show isotopic anomalies of its own. If, however, xenon *H* was made by fission in the meteorite, its carrier ought to show evidence of fission, such as other fission products. In collaboration with B. Srinivasan, who is now at Washington State University, we undertook the project.

As so often in our work, we did the right experiment for the wrong reasons. On the mistaken hunch that xenon *H* is located in the mineral pentlandite $[(Fe,Ni)_9S_8]$ we dissolved a sample of the Allende carbonaceous chondrite in hydrochloric acid and hydrofluoric acid, which supposedly do not dissolve pentlandite. A black residue remained, representing about .5 percent of the mass of the meteorite, and it indeed contained most of the xenon *H*. But pentlandite was only a small and, as it turned out, irrelevant part of the residue. Ultimately the residue proved to consist mostly of amorphous carbon, chromite (Cr_2FeO_4) and spinel.

Annoyingly, the residue also con-



ANOTHER TYPE of anomalous xenon emerged from the stepped heating of a residue of the Murchison meteorite. Srinivasan and one of the authors (Anders) had treated it with reagents that oxidize organic polymers. The intent was to remove primordial xenon. The first two fractions of gas still were close to the line representing mixtures of primordial xenon and the xenon Reynolds and Turner discovered. Then the fractions veered away from the line. The aberrant fractions proved to contain a type of xenon (*s* xenon) consisting mainly of middle isotopes.



ISOTOPIC COMPOSITION of *s* xenon (*colored bars*) differs from that of solar xenon (*light gray bars*) but matches that due to the *s* process (slow neutron capture) in red-giant stars (*dark gray bars*). *S*-process pattern was calculated by D. D. Clayton and R. A. Ward of Rice University.

tained abundant planetary xenon, which thoroughly masked the xenon *H*. We now made a second lucky mistake. Still thinking that at least one of the xenon components was trapped in a sulfide, we etched the residue with nitric acid, which is known to dissolve all sulfides. Sure enough, the primordial xenon in the residue was lost, whereas the xenon *HL* stayed behind, in much purer form than ever before. Actually, however, none of the xenon is trapped in a sulfide; the experiment succeeded for quite a different reason. The primordial xenon is trapped at the surface of grains and is lost when nitric acid etches the surface away. Xenon *HL* is trapped inside the grains and stays behind.

At this point, then, we had managed to concentrate the carrier (or carriers) of xenon *HL* in a residue representing a minute fraction of the mass of the meteorite. Now we could look for isotopic anomalies due either to stellar nucleosynthesis or to fission. The first tests were negative: both carbon and osmium were normal. Then Robert Clayton and Mark H. Thiemens at Chicago and Urs Frick and Pepin at Minnesota tested oxygen and nitrogen. Both were distinctly anomalous. Since the residue still contained a mixture of carbon, chromite and spinel, however, it remained unclear whether the anomalies were associated with the actual xenon carriers. The mixture of minerals was analogous to a set of nested Russian dolls. If xenon *HL* was in the innermost doll, it would do no good at all to measure the isotopic composition of other elements in the outer dolls. The problem was to find the innermost doll and make all measurements on that.

Next spinel was eliminated, on the ground that pure spinel extracted from the residue had almost no trapped gases in it. Carbon and chromite remained. But a detailed study of those minerals, done in part by Frick and Ulrich Ott of Berkeley and Sherwood Chang of the Ames Research Center of the National Aeronautics and Space Administration, showed that the major part of the carbon and chromite is poor in trapped gases. Nearly all the xenon *HL* is thus in a small part of the carbon (and possibly chromite), which may or may not be the innermost Russian doll.

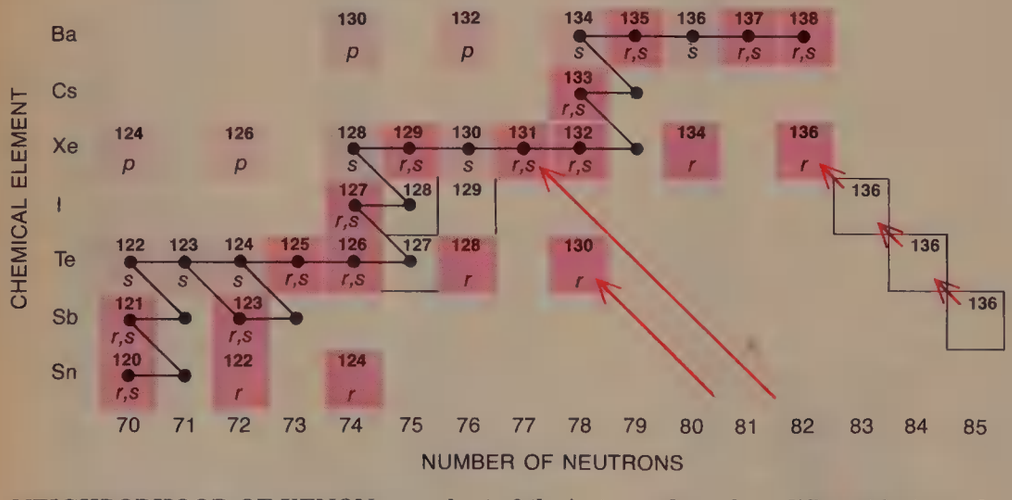
Last year P. K. Swart, M. M. Grady and C. T. Pillinger of the University of Cambridge determined the isotopic composition of the carbon. They found that the ratio of carbon 12 to carbon 13 was 93, a number within the range for terrestrial carbon (88 through 93). Taken at face value this result supports a local origin for the carbon and hence the xenon, since a ratio of 93 is quite uncommon in the galaxy. Values in stars range from less than 10 to several hundred, and the average for interstellar clouds is 62.

Troubles for Fission

Tests for fission were inconclusive, and so a stalemate persisted for more than 10 years, until two crucial experiments were done at last in 1982. In the first one we, working with I. P. Wright, S. J. Norris and Pillinger of Cambridge, measured the isotopic composition of the nitrogen in two residues from the Murchison carbonaceous chondrite, one of which, the more fine-grained one, was enriched nearly 600-fold in xenon *HL*. This residue proved to contain very unusual nitrogen, enriched in nitrogen 14 by nearly 30 percent. Such an anomaly could have resulted only from stellar nucleosynthesis, not from fission, and so if the fine-grained sample is indeed the innermost Russian doll, xenon *HL* must have also been made in a star. We are left with the puzzle of why the carbon in the sample is so perfectly normal. One marvels at nature's sense of humor in choosing carbon of terrestrial isotopic composition as a package for highly anomalous nitrogen and xenon.

In the second experiment we, working with G. W. Lugmair and Tadashi Shimamura of the University of California at San Diego, tested the Allende meteorite for isotopic anomalies in barium, neodymium and samarium. Any process that made xenon *HL* should also have made comparable amounts of at least some isotopes of these neighboring elements, in proportions that would vary from one process to another. Surprisingly, we found no anomalies whatever (except for enrichments of neodymium 142 and 143, both attributable to radioactive decay). That finding seems at first an equal embarrassment for both hypotheses about the origin of xenon *H*, stellar nucleosynthesis and fission inside the meteorite. For nucleosynthesis, chemistry offers a possible way out of this predicament. Barium, neodymium and samarium are quite involatile and thus would be among the first elements to condense in the expanding, cooling gas expelled by a supernova. Presumably they would condense on oxide minerals such as perovskite ($CaTiO_3$). Xenon, being a noble gas, would condense much later, at lower temperatures and on a different substrate, presumably carbon.

No such excuse is available for the fission hypothesis, and so we are left with stellar nucleosynthesis. One possibility is some variant of the *r* process during explosive carbon "burning," which may occur in a supernova when a shock wave passes through a carbon-rich zone of the dying star. Calculations by Dieter Heymann and Marlene Dziczkaniec of Rice suggest that the neutrons and gamma rays (high-energy photons) released at a temperature of two billion degrees rework the isotopic composition of heavy elements in the zone so that xenon emerges in the right isotopic proportions. Another possibility is a faster *r* process that builds nuclei all the way up to short-lived superheavy elements of mass near 280, which then fission into lighter elements including xenon. Calculations by E. P. Steinberg and B. D. Wilkins of the Argonne Na-



NEIGHBORHOOD OF XENON on a chart of the isotopes shows how different isotopes derive from different processes: the *s* process (*s*), the *r* process (*r*) and the *p* process (*p*). Each involves characteristic moves on this nuclear chessboard. In the *s* process stable atomic nuclei slowly capture neutrons (*horizontal black lines*). Whenever a capture makes a nucleus unstable, that is, radioactive (*white boxes*), a beta decay (*diagonal black line*) intervenes: the nucleus emits an electron, converting one of its neutrons into a proton. The *s* process cannot make the two lightest and the two heaviest isotopes of xenon; thus it fits the pattern of the xenon discovered by Srinivasan and Anders. In the *r* process—rapid neutron capture—neutron-rich nuclei (mostly off scale to the lower right) are built up so fast that even the short-lived radioactive ones have little chance to decay. When the supply of neutrons is gone, a series of beta decays converts them into stable isotopes. One series, leading to xenon 136, is shown (*string of diagonal colored arrows*). The radioactive nuclei made by fission decay along similar paths. Some stable isotopes (*light colored boxes*) cannot be made by beta decay. They are "shielded": a stable nucleus richer in neutrons (*dark colored boxes*) ends the sequence. The shielded isotopes form only by the *s* process or the *p* process: proton capture or the ejection of neutrons by gamma rays.

tional Laboratory show that the fission would help to explain the high solar-system abundance of the rare-earth elements terbium, dysprosium, holmium and erbium. It also accounts for the xenon *H* data rather well.

Whether or not short-lived superheavy elements exist may ultimately be decided by efforts to synthesize them at the major nuclear research centers at Berkeley, Dubna in the U.S.S.R. and Darmstadt in Germany. Prospects for a long-lived superheavy element have dimmed greatly, however, as a result of the latest work on xenon *H*.

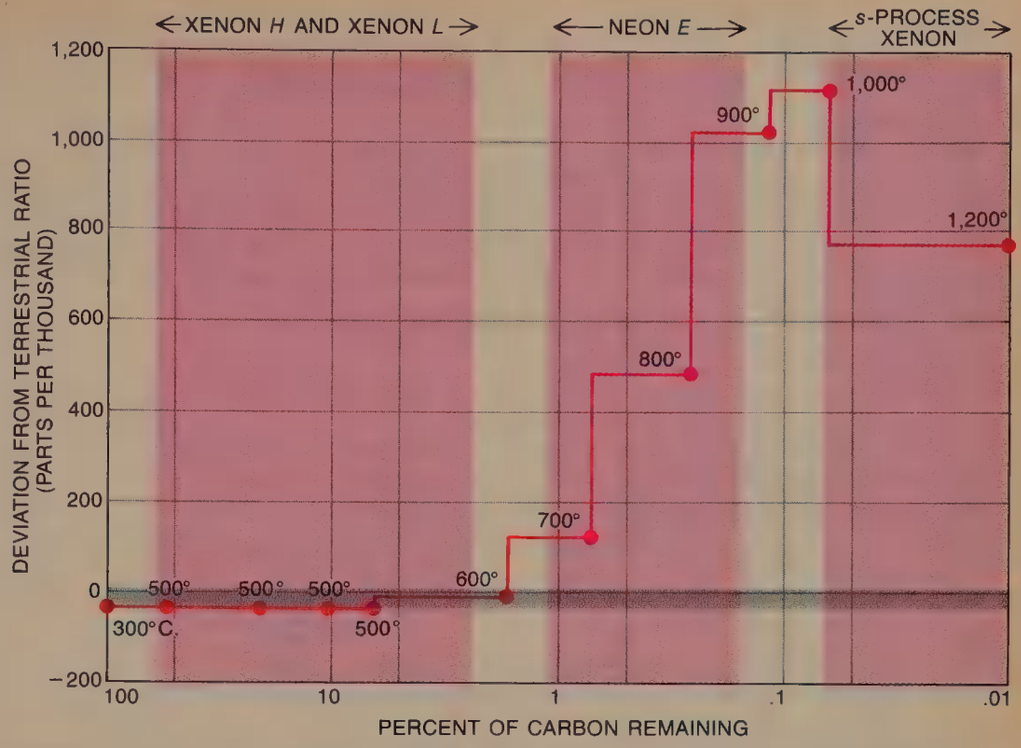
Dust Grains from Red Giants?

While struggling with xenon *HL* we accidently discovered a third anomalous type of xenon. It is rarer than the others and also better hidden, but its isotopic composition points quite plainly to a specific place of origin.

Again we did the right experiment for the wrong reason. Srinivasan and one of us (Anders) were trying to characterize the xenon *HL* in the Murchison meteorite in the hope that it might differ from the xenon in Allende in its ratio of xenon *L* to xenon *H*. (Murchison is a C2 carbonaceous chondrite, Allende is C3.) To our disappointment the etched residues from Murchison still contained a lot of primordial xenon. Presumably the nitric acid treatment had failed to remove all the organic polymer. (Organic polymer contains primordial gases and is present in C2 chondrites but not in C3.) We therefore treated the residues with NaOCl and Na_2O_2, compounds that oxidize organic polymers. Then we examined the samples by stepped heating.

The first two data points showed what we were looking for. Primordial xenon came off at 800 degrees C., followed at 1,000 degrees by xenon *HL* purer than any Murchison residue had ever yielded before. The next four points were surprises. Beginning at 1,100 degrees they veered sharply away from the line on our chart representing mixtures of primordial xenon and xenon *HL*. Apparently the sample contained a new xenon component lying at or beyond the 1,600-degree point. It constitutes less than one part in 10^4 of the total xenon, so that we had missed it in all previous experiments.

The new component proved to be enriched mainly in the even-numbered isotopes xenon 128, 130 and 132 and grossly deficient in the lightest and the heaviest isotopes xenon 124, 126, 134 and 136. This pattern was highly characteristic: it clearly pointed to the *s* process. By coincidence the theoretical pattern for *s*-process xenon had just been calculated by Donald Clayton and R. A. Ward; it matched our new component perfectly.



COMBUSTION OF A CARBON RESIDUE from the Murchison meteorite shows that it hides three exotic types of carbon; each one carries anomalous neon or xenon. The sample was heated in steps in the presence of oxygen; in each resulting fraction of carbon dioxide the ratio of carbon 13 to carbon 12 was determined. The first fractions came from the carrier of xenon *H* and xenon *L*. By chance their carbon had isotopic proportions like those of carbon on the earth (*gray band*). Later fractions came from the carriers of neon *E* and *s* xenon. The experiment was done by P. K. Swart, M. M. Grady and C. T. Pillinger of the University of Cambridge.

The *s* process, like the *r* process, builds heavy elements by neutron capture, but much more slowly: intervals of years rather than milliseconds intervene between successive neutron captures by any one nucleus. Hence any short-lived radioactive nuclei formed will tend to decay before capturing another neutron. The buildup of nuclei thus follows an "*s*-process path" involving neutron captures punctuated by decays. Take tellurium 122, a stable isotope. The *s* process converts it sequentially into tellurium 123, 124, 125 and 126, all of which are stable, or nonradioactive. Then comes tellurium 127, a radioactive isotope. It decays into iodine 127. (Specifically, a neutron in the nucleus of tellurium 127 emits a beta particle, or energetic electron, and is thereby converted into a proton.) The capture of another neutron yields iodine 128, which is radioactive; then xenon 128 through xenon 132; then xenon 133, which is radioactive; then cesium 133, and so on. Note that the *s* process does not make the lightest and heaviest xenon isotopes, the very ones that are missing in the new meteorite component.

The *s* process occurs in red giants: stars that have exhausted the hydrogen in their core and shine by thermonuclear reactions that convert helium 4 into carbon 12. They also burn the hydrogen that remains in a gaseous shell farther outward. The site of the *s* process is below the hydrogen-burning shell, where neutrons are freed by reactions such as the one fusing neon 22 and helium 4 into magnesium 25, allowing a neutron to escape. The neutrons then are captured by iron 56 and other heavy nuclei; thus the *s* process begins.

At one time, therefore, the *s*-process xenon, *s* xenon for short, must have been inside a red-giant star. When and how did it escape? Perhaps at the red-giant stage; red-giant stars are known to emit a wind of particles by which they lose mass at rates as great as 10^{-7} times the mass of the sun per year. Perhaps at some later time, when a red-giant star evolves into a nova or supernova. Matter is ejected at all the stages of a red giant's evolution and cools by expansion and radiation until its temperature is sufficiently low (less than 2,000 degrees K.) for grains to condense. Stable isotopes of xenon and other heavy elements, once made by the *s* process, undergo no further change in a red giant until the star becomes a supernova and temperatures rise above two billion degrees. In contrast, the lighter chemical elements such as carbon, nitrogen, lithium, helium and neon continue to react as the star evolves and heats up. Therefore we turned to the lighter elements. Their isotopic composition might provide some clues to the precise stage at which the grains bearing *s* xenon were ejected.

Heavy Carbon

We soon found that the *s* xenon in meteorites is in carbon. We treated our residue of the Murchison meteorite with perchloric acid, which oxidizes carbon

into carbon dioxide while leaving most minerals unaffected. The *s* xenon disappeared.

There was reason to expect this carbon to be isotopically anomalous, because red giants and their successor stars have ratios of carbon 12 to carbon 13 that vary more than a hundredfold. The problem was how to resolve this exotic carbon from the much larger amount of ordinary carbon that presumably was still in the sample. We used the fine-grained Murchison residue that had been analyzed for nitrogen. It had been demineralized with acids, etched with alkaline oxidizing agents to remove the polymer and finally ultrafiltered to remove material coarser than one micron. It was analyzed by Swart, Grady and Pillinger.

To resolve the various carbon components they used a stepped-combustion technique invented by David J. Des Marais of the Ames Research Center, in which a sample is heated to progressively higher temperatures with oxygen and the carbon dioxide formed at each step is analyzed in a mass spectrometer. The different types of carbon in the sample burn off according to their grain size and crystallinity and with luck can be analyzed one by one.

The major part of the carbon proved to belong to the carrier of xenon *HL* and had its characteristic composition, some 3 to 4 percent lighter than standard carbon. Above 600 degrees C., however, the carbon got increasingly heavy, reaching a 110 percent enrichment in carbon 13 at 900 to 1,000 degrees. Such an extreme composition is unprecedented. Apparently a new, isotopically heavy type of carbon is present in the meteorite. If the extremest enrichment of 110 percent represents this heavy carbon in pure form, our Murchison residue contains .45 percent of heavy carbon and the bulk Murchison meteorite contains about five parts per million.

Actually the heavy carbon is probably a mixture of two types of carbon: the carrier of *s* xenon and that of neon *E*(*L*). The carrier of xenon *HL* is also carbon. This gives a total of three types of exotic, presumably presolar, carbon in primitive meteorites.

A Tracer of Interstellar Molecules

The last trail we shall follow involves deuterium (hydrogen 2) and leads to a fourth type of exotic carbon. The first hint came in 1953. Giovanni Boato, working in Harold C. Urey's laboratory at Chicago, found that several carbonaceous chondrites were enriched in deuterium by up to 31 percent (with respect to ocean water). For many years no attempt was made to trace the enrichment to one of the principal chemical forms of hydrogen in carbonaceous chondrites: clay minerals containing OH groups and organic matter containing CH groups (mainly the organic polymer mentioned above).

The key experiment was done at last in 1979 by Y. Kolodny, J. F. Kerridge and I. R. Kaplan of the University of California at Los Angeles. Stimulated by the work of F. Robert and his colleagues at the Saclay Nuclear Research Center (CENS), they analyzed samples of carbonaceous chondrites by stepped heating before and after the organic matter in the samples had been burned off in a plasma (an ionized gas) of oxygen; they inferred the deuterium content of the organic matter from the difference between the two analyses. The deuterium enrichment turned out to be localized in the organic matter, with excesses ranging up to 160 percent. Later measurements by Robert, Richard H. Becker, Jiyoung K. Yang and Samuel Epstein of the California Institute of Technology extended the range to 310 percent and showed that of the various types of organic matter it is the polymer that shows the greatest enrichment.

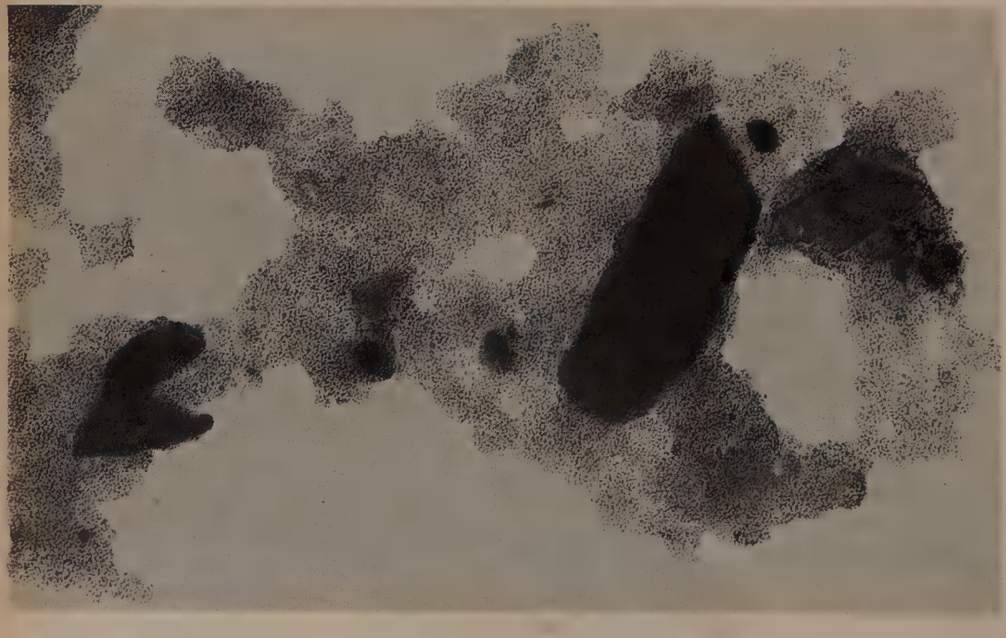
Deuterium is the most fractionation-prone of all isotopes, having twice the mass of its sister isotope hydrogen 1. An enrichment of 310 percent, or a factor of 4.1, is nonetheless hard to explain. In fact, the enrichment is even greater, because seawater itself is enriched eightfold with respect to galactic hydrogen. Thus the overall enrichment of the meteoritic polymer is 32-fold.

In principle the enrichment could be produced by isotopic exchange reactions such as the one that transfers a deuterium atom (D) from molecular hydrogen to methane (CH_4): $CH_4 + HD \leftrightharpoons CH_3D + H_2$. According to theoretical calculations, this reaction enriches the deuterium content of methane 32-fold at a temperature of about 130 degrees K. Similar enrichments occur when the organic matter forms from small molecules. The catch is that reaction rates at 130 degrees are ludicrously low. The half-time for the formation of methane from carbon monoxide and molecular hydrogen is 10^{30} years.

J. Geiss of Bern and Hubert Reeves of Saclay have developed a better explanation from a suggestion they first made in 1972. Molecules in interstellar space such as HCHO and HCN are enriched in deuterium by factors of up of 10^5, according to their radio spectra. Perhaps the meteorites contain a trace of such interstellar molecules.

Ion-Molecule Reactions

Why are interstellar molecules enriched in deuterium? The question was answered in 1973 by William D. Watson of the University of Illinois at Urbana-Champaign. It was known at the time that interstellar molecules form mainly by a series of ion-molecule reactions, which are fast even at the lowest temperatures. Typically a carbon ion made by cosmic rays reacts with a hydrogen molecule to form the molecular ion CH_2^+. The molecular ion then reacts with other hydrogen molecules to form a more complex ion. Finally it reacts with a free electron to form a stable, neutral molecule such as CH_4. Watson noted that the chemical bond between carbon and deuterium is slightly stronger than the bond between carbon and hydrogen 1, causing deuterium to concentrate in ions during exchange reactions. For the same reason ions shedding hydrogen will tend to lose hydrogen 1 rather than deuterium, causing deuterium to concentrate still further in the molecular ion and ul-

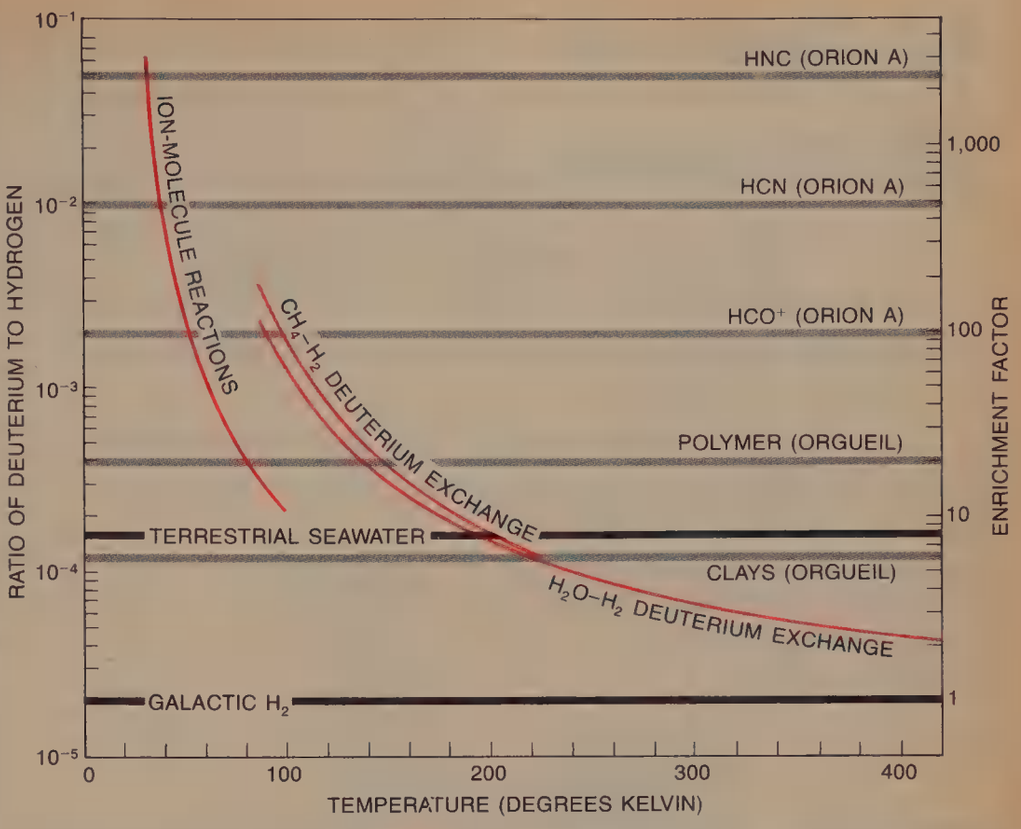


ELECTRON MICROGRAPH of a Murchison sample containing the three exotic types of carbon was made by Mitsuo Ohtsuki of the University of Chicago. Only one of the types can be pinpointed: the fine-grained material carries xenon *H* and xenon *L*. The other two types make up a few percent of the larger, darker grains, which range up to a micrometer in size.

timately the neutral molecule. The extent of the enrichment depends on the particular molecule but rises with falling temperature.

The deuterium-rich matter in meteorites is present as a complex organic polymer, not as pristine small molecules such as those detected by radio astronomers. This is not suprising, since these molecules are too volatile and also too reactive to persist unchanged. The carbon in the polymer is isotopically unexceptional (1.6 percent lighter than standard carbon), suggesting either that the carbon in interstellar molecules is unexceptional or more probably that the interstellar carbon is greatly diluted by "local" (primordial) carbon.

The local polymer, in any case, has a complex structure, consisting of aromatic ring systems with up to four fused rings joined by hydrocarbon bridges $[(CH_2)_n]$. That tells us little about its origin; coal formed from plant matter is similar. Apparently many kinds of organic matter form aromatic polymers on sustained heating. The deuterium-rich, interstellar part of the polymer need not have the same structure, because many interstellar molecules, including HNC, $H(C\equiv C)_nCN$ and HCHO, are quite reactive and may have polymerized on the surface of grains at low temperatures. Perhaps the difference will make it possible to separate the local and exotic polymers by chemical means.



ANOMALOUS DEUTERIUM reveals a fourth exotic type of carbon. In the chart the ratio of deuterium to light hydrogen (that is, hydrogen 2 to hydrogen 1) is compared for a number of sources of hydrogen. The ratio in our galaxy serves as a standard. The earth's seawater proves to be enriched by a factor of eight. Clays from the Orgueil meteorite are enriched slightly less than that; carbon polymer from Orgueil is enriched markedly more. Molecules in interstellar clouds such as Orion A are enriched up to 100,000-fold. In principle, reactions that transfer deuterium from hydrogen molecules to other molecules (*light color*) could cause the polymer's enrichment; the reactions, however, are slow. A more likely possibility is that the polymer includes a trace of interstellar molecules, whose enrichment comes when molecular ions such as CH_2^+ react with hydrogen molecules (*dark color*). The illustration was devised by J. Geiss of the University of Bern and Hubert Reeves of the Saclay Nuclear Research Center (CENS).

Solar-System Carbon

In the course of this article four types of exotic carbon in primitive meteorites have been discussed; three are forms of elemental carbon carrying neon *E*(*L*), xenon *HL* and *s* xenon, and the fourth is a polymer enriched in deuterium. Each is hidden in local carbon. How did carbon end up in such a wealth of chemical states?

We can try to answer the question in terms of thermodynamics, which predicts the carbon compounds that should have formed in the solar nebula at various temperatures and pressures. Let us therefore see what happens when a solar gas cools from high temperatures at a pressure of 10^{-5} atmosphere, a reasonable value for the region of the future asteroid belt. (The model we shall present is based on work Ryoichi Hayatsu and one of us, Anders, have done with M. H. Studier of Argonne.) Initially carbon is present mainly as gaseous carbon monoxide, the dominant form of carbon in interstellar space. It remains in that form as the gas contracts and heats up to form the solar nebula. On cooling it ought to hydrogenate into methane. In the absence of catalysts, however, the rate of reaction is very low. The same is true of the reaction that becomes feasible at somewhat lower temperatures

TYPE OF CARBON	CLASS OF METEORITES	CONCENTRATION OF CARBON (PARTS PER MILLION)	NOBLE GASES OR OTHER MARKERS	ORIGIN	RATIO OF CARBON 12 TO CARBON 13
ORGANIC	C1, C2	ABOUT 30,000	PRIMORDIAL	LOCAL ($CO + H_2$)	91
ELEMENTAL	C3	2,000	PRIMORDIAL	LOCAL ($2CO \rightarrow C + CO_2$)	91
CARBONATE	C1 SOME IN C2	2,000 OR LESS		LOCAL ($CO_2 + MgO$)	83-85
CARBON-ALPHA	C1, C2 SOME IN C3	5 OR LESS	NEON *E*	NOVA?	80 OR LESS
CARBON-BETA	C1, C2 SOME IN C3	5 OR LESS	*s*-PROCESS XENON	RED GIANT	42 OR LESS
CARBON-DELTA	C1, C2, C3	200	XENON *H*, XENON *L*, AND LIGHT NITROGEN	SUPERNOVA	92
ORGANIC	C1, C2	3,000 OR LESS	DEUTERIUM	INTERSTELLAR MOLECULAR CLOUD	?

TYPES OF CARBON in carbonaceous chondrites are summarized. Three types are local: their properties can be explained by processes in the early solar system. Four types are exotic: their interstellar origin is revealed by the isotopic composition of the carbon or by isotopic anomalies in "markers" that they carry. Each exotic type is discussed in the text of this article.

To preserve your copies of SCIENTIFIC AMERICAN

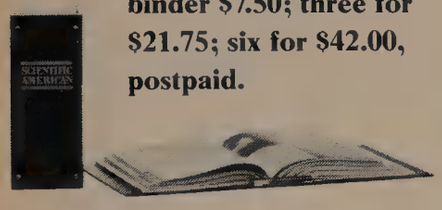
A choice of handsome and durable library files—or binders—for your copies of SCIENTIFIC AMERICAN. Both styles bound in dark green library fabric stamped in gold leaf.



Files **Each file holds 12 issues. Price per file $5.95; three for $17.00; six for $30.00, postpaid.**

(Add $2.50 each outside U.S.A.)

Binders **Each binder holds 12 issues. Issues open flat. Price per binder $7.50; three for $21.75; six for $42.00, postpaid.**



(Add $2.50 each outside U.S.A.)

To: Jesse Jones Box Corp.
P.O. Box 5120
Philadelphia, PA 19141

Send me_____
SCIENTIFIC AMERICAN
☐ Files ☐ Binders
For issues dated ☐ through 1982
☐ 1983 or later.
I enclose my check or money order for $_____ *(U.S. funds only).*

Name _____
(please print)

Address _____

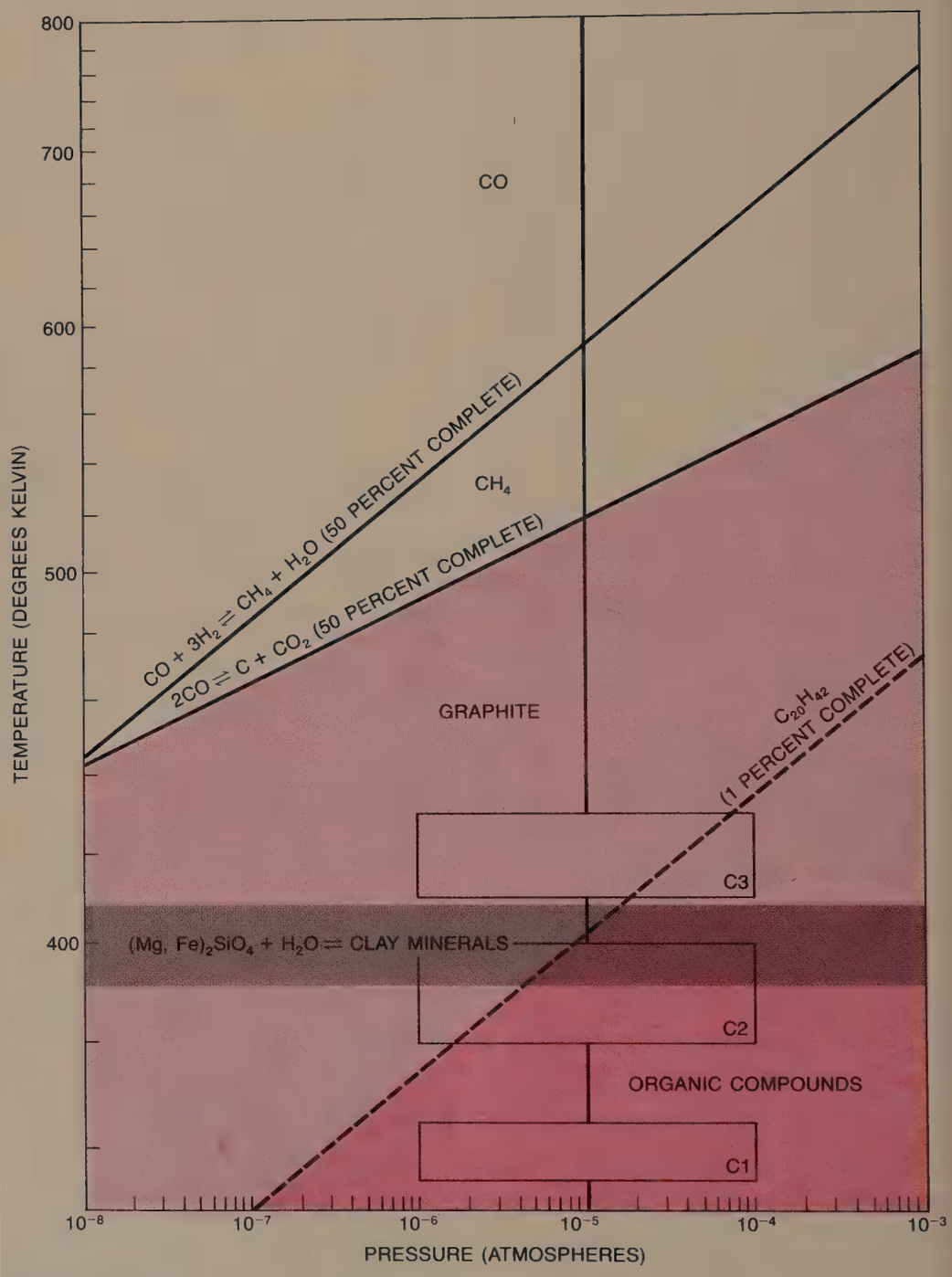
City _____

State _____ Zip _____

NOTE: Satisfaction guaranteed or money refunded. Allow four to six weeks for delivery.

and converts carbon monoxide into carbon dioxide and elemental carbon.

Below 400 degrees K. the situation changes drastically. Clay minerals and magnetite form, both of which are excellent catalysts for the hydrogenation of carbon monoxide, but at these lower temperatures methane, still the most stable hydrogenation product, is no longer the only possible one. Higher hydrocarbons ($C_{20}H_{42}$, for example) can form along with many other organic compounds. Actually they do, as the reaction pathway favors the growth of long hydrocarbon chains. The reaction is essentially the industrial Fischer-Tropsch process: the conversion of carbon monoxide and hydrogen into hydrocarbons, alcohols and other compounds in the presence of a catalyst.



CHEMISTRY OF CARBON in the solar nebula explains the various states of the local carbon in carbonaceous chondrites. At high temperatures carbon monoxide (CO) is the dominant form of carbon in the nebula. Then the nebula cools. The CO ought to be converted into methane (CH_4); a sloping line shows where the reaction is half complete. In the absence of catalysts, however, the conversion is slow, and so most of the CO survives. Next carbon dioxide (CO_2) ought to appear along with elemental carbon (C) in the form of graphite or amorphous carbon grains. Again the conversion is slow. Near 400 degrees Kelvin a dramatic change begins. Silicates react with water, forming clay minerals that catalyze the formation of complex organic compounds. A line charts the combinations of temperature and pressure at which the conversion of carbon monoxide and hydrogen into $C_{20}H_{42}$, a typical complex molecule, is 1 percent complete. Three rectangles in the illustration show the ranges of temperature and pressure in which carbonaceous chondrites presumably formed. C3 chondrites such as Allende formed at temperatures greater than 400 degrees K.; they contain small amounts of mostly amorphous carbon. C1 chondrites such as Orgueil and C2 chondrites such as Murchison formed at lower temperatures. They contain organic compounds and also the clay minerals that were catalysts.

The industrial process was developed in 1923. The meteoritic hydrocarbons show the signature of the reaction: predominance of straight carbon chains.

The solar-nebula sequence explains the diversity of the local carbon found in carbonaceous chondrites. C3 chondrites, which according to the calculations of J. W. Larimer of Arizona State University formed at temperatures between 410 and 430 degrees K., should contain mainly amorphous carbon but no organic compounds, and that is the case. C1 and C2 chondrites, which formed at temperatures of less than 400 degrees, should contain mostly organic carbon, and they do. The two forms of organic carbon found in C1 and C2 chondrites (soluble compounds and polymer insoluble in standard organic solvents) may represent different lengths of contact between the organic matter and the mineral catalyst. Fischer-Tropsch reactions in the laboratory at first yield only soluble compounds, but after a period of six months these compounds are transformed into complex insoluble matter resembling the meteoritic polymer. The small amount of carbonate in meteorites may have formed when oxides of magnesium, calcium and iron (by-products of the formation of clay minerals) reacted with carbon dioxide (a Fischer-Tropsch by-product).

For presolar carbon one further possibility is available: direct condensation of graphite grains at high temperature. Thermodynamic calculations show that such condensation is feasible only in a gas whose ratio of carbon to oxygen is greater than .9, well above the solar ratio of .6. Red giants and their successor stars tend to have such ratios. In addition there are indications that the dust shells surrounding red giants do contain graphite grains. Such grains ought to be better crystallized and thus more resistant to chemical change than the amorphous carbon that forms at lower temperatures from carbon monoxide. This may well be the reason we were able to purify presolar carbon by techniques of partial combustion and chemical oxidation. In the process, however, we probably lost more reactive types of carbon.

Given only quanta of light and the laws of physics, astronomers and astrophysicists have inferred the nature and inner workings of a wealth of astronomical objects, including red giants, novas, supernovas and interstellar clouds. Now new clues are becoming available: bits of tangible matter from the objects themselves, each containing a rich record of stellar nucleosynthesis and interstellar chemistry. The potential yield of knowledge is great, since laboratory measurements can reveal much detail that is inaccessible to astronomical techniques. Fifteen years ago there seemed no prospect of ever studying stardust. Today the stardust is in hand.

How to tame your data.



To soothe the savage information beast, a businessperson could use the IBM Personal Computer XT.

Because with XT's 10-million-character fixed disk drive and IBM data management software *specifically* designed to complement the hardware, you can whip thousands of names and numbers into more manageable shape. (Helping you get a better shot at the lion's share.)

Use IBM PFS:FILE* to generate a "form" on the screen. Customize it by putting pertinent data in the blank spaces provided.

Then use IBM PFS:REPORT to sort, organize, search, update, store and print the facts with ease.

To learn more about how the IBM Personal Computer XT can help you more efficiently handle many high-volume applications, visit your authorized IBM Personal Computer retail dealer. IBM®



The IBM Personal Computer
A tool for modern times

For more information on where to buy the IBM Personal Computer, call **800-447-4700.** In Alaska or Hawaii, **800-447-0890.**
* PFS is a registered trademark of Software Publishing Corporation

The Chemical Defenses of Termites

Soft-bodied and blind, termites are subject to predation. In defense termite soldiers attack intruders with an array of sophisticated irritants, toxins, anticoagulants and glues

by Glenn D. Prestwich

For more than 100 million years the insect world has had a chemical arms race. Insect predators attempt to subdue their prey with toxic venoms or to lure them with seductive perfumes. Insect prey species counter by ejecting irritating, sticky, hot or poisonous substances or by being inedible. Unique among the chemically defended insects is the termite caste of soldiers, whose heads and bodies are so thoroughly modified into weapons that they can neither feed themselves nor reproduce. No other insect species, not even the army ants, have such a totally specialized and dependent full-time army. Nor does any other insect order exhibit such a diverse arsenal of chemical weapons and delivery systems.

Termites evolved some 150 million years ago from an ancestral stock resembling the modern roaches. Today there are more than 2,000 species of termites around the world, 95 percent of them in the Tropics of the Old World and the New. They make up the order of insects known as the Isoptera ("equal wings"), which are the most primitive of the social insects.

Termites, often incorrectly called "white ants," evolved quite separately from the social ants, bees and wasps of the Order Hymenoptera. Termite social organization evolved around the need for food sharing to exchange the bacterial and protozoan symbionts necessary to digest cellulose. Termites have rigidly structured societies in which morphologically specialized individuals execute specific tasks: the king and queen reproduce, the workers forage, build the shelter and care for the young, fertile winged pairs fly off to form new colonies and the soldiers defend. Communication among individuals of the colony is based on the exchange of chemical signals, both by smell (olfaction, using the antennae) and by taste (contact chemoreception). Building, food finding, nestmate recognition, trail following, alarm and defense all involve specific chemical cues.

Termite soldiers can be sterile males or sterile females, and their heads are much different in shape and size from those of termite workers. The soldiers' primary role in the activities of a termite colony is to protect it from intruders. This they do continuously from maturity to death. A dead soldier is to the colony what dead epidermal cells are to the human body; the loss increases the probability of genetic transfer to the next generation. As a result termite soldiers are walking weapons and can execute their defensive role by tactics that include biting, snapping, hole plugging, squirting, oozing, daubing and even self-destructive defecation.

Here I should like to discuss certain termite soldiers representative of both the "lower" and "higher" termites. I shall be placing particular emphasis on those genera of both groups that have supplemented their mechanical weapons (enlarged mandibles) with chemical ones. Many different kinds of disagreeable secretion can be applied to predatory intruders in many different ways. I shall also be showing how the analysis of the specific chemical composition of these different secretions helps in tracing the evolutionary pathways followed by different termite lines, in some instances for as long as 70 million years.

Although the weaponry of termite soldiers can be subdivided into the mechanical and the chemical, the full range of defensive tactics is even broader. For example, the soldiers of some species puncture the cuticle of an intruding insect with their powerful mandibles and then coat the wound with an anticoagulant secretion from their frontal gland. Neither the wound nor the applied secretion would in itself be fatal to the intruder, but the combination of a gashed cuticle and a chemical that either prevents clotting or is internally toxic means the wounded insect will eventually succumb.

At the same time at least two other defensive tactics depend on neither mandibles nor secretions. The head of the *Cryptotermes* soldier is shaped something like a drain plug, and the soldiers' reaction to an alarm is to disperse to each of the colony's narrow entrances and literally seal off the cylindrical holes with their pluglike heads. Colonies of *Anoplotermes* do not even have a soldier caste, but each worker in the colony has an abdomen ringed with a specialized constricting muscle. When the worker is confronted by an intruder, it contracts the muscle, rupturing its abdomen and bathing the attacker with a mixture of feces and other gut contents.

What has most attracted the interest of entomologists in recent years is the remarkable variety of termite defensive secretions. Two investigators who have studied the morphology and glandular structure of termite soldiers' chemical weapons are André Quennedey of the University of Dijon and Jean Deligne of the Free University of Brussels. They recognize three main methods of chemical defense. The first is biting, with the simultaneous introduction of an oily or toxic material into the wound. The second is daubing, with a contact poison being applied to the cuticle of an attacker with an enlarged labrum, or upper lip, resembling a paintbrush. The third is "glue squirting": the soldier sprays at the aggressor an irritating, viscous entangling agent.

The first of the three methods has evolved independently a number of times in the family of the higher termites, the Termitidae, and also in one of the six families of the lower termites, the underground-dwelling Rhinotermitidae. More than 40 "biting/injecting" species have now been studied in my laboratory at the State University of New York at Stony Brook and by investigators in the Chemical Entomology Unit at the University of Southampton. Something of the variety in chemistry and morphology that has evolved along these lines can be illustrated by considering three biting/injecting genera.

Colony defense in one of these genera, the higher-termite genus *Macrotermes*, is the task of a soldier caste of sterile females with a dual membership. Small



SOLDIER TERMITE of the species *Nasutitermes corniger* is magnified 50 times in this scanning electron micrograph. *N. corniger* is one of the more than 500 species in the higher termite subfamily, the Nasutitermitinae. The subfamily is named for the elongated nasus, or snout, extending from the soldiers' forehead. With this bazooka-like proboscis the nasute termites squirt a gluelike substance on intruders.



TOP OF THE HEAD of a soldier of the same species of termite appears in this scanning electron micrograph magnified 70 diameters.



TIP OF THE SNOUT of a soldier of the same termite species is further magnified 1,100 diameters in this scanning electron micrograph.

Macrotermes soldiers escort the large sterile male workers as they gather food and repair the colony's mound. Large soldiers guard the inner hive, where the king and queen and the larvae are sheltered. In soldiers of both sizes the frontal gland of the head secretes an oily substance and exudes it through the fontanelle, a glandular opening in the head cuticle.

The secretion is chemically the same for both kinds of soldiers, but the large soldiers carry 500 times more of it than the small ones. Indeed, it amounts to almost 10 percent of the large soldiers' dry weight. The secretion is a hydrocarbon similar to mineral oil or paraffin: it consists of long-chain alkanes and alkenes [*see illustration on opposite page*]. The chains, made up of two-carbon acetate units linked together, are from 21 to 35 carbon atoms long; the exact composition of the molecule varies with the species of termite and the location of the colony.

The commonest raiders of termite colonies are ants. When *Macrotermes* soldiers fight an ant, they bite furiously with their mandibles in an effort to snip off the intruder's legs. The heat of the activity turns the waxy secretion in the frontal gland into a free-flowing oil that oozes from the fontanelle, runs down the soldier's rostrum, or forehead, and spreads out onto the labrum overlying the mandibles. Thus every successful bite applies a liberal coating of the oil to the ant's punctured cuticle. In the absence of a puncture the nontoxic hydrocarbon is harmless. When it is applied to injured cuticle, however, it appears to soften the punctured area. Hence the application interferes not only with the coagulation of the ant's hemolymph but also with the resclerotization, or natural repair, of cuticle damage.

The second biting/injecting genus among the higher termites is *Cubitermes,* one of the soil-eating termite genera found in Africa. The soldiers have saberlike mandibles and squarish orange heads. They too secrete a hydrocarbon. The *Cubitermes* secretion, however, is not a simple straight-chain molecule but belongs to the family of terpenes [*see illustration on opposite page*]. Terpene molecules are grouped according to the number of five-carbon isoprene units in them. Therefore two-isoprene units (10 carbons) are referred to as monoterpenes, four-isoprene units (20 carbons) as diterpenes and so on. Diterpenes are secreted by very few insects. In fact, except for some of the higher termites no insects have been shown to be capable of synthesizing them. *Cubitermes* soldiers, however, are in the diterpene manufacturing business on a large scale. As postdoctoral associates of Jerrold Meinwald at Cornell University, David Wierner and I identified three new diterpenes secreted by this termite genus. Of the more than 16 diterpenes found among seven species of *Cubitermes* I now know at least five that are unique to these termites. No other invertebrates, vertebrates or even plants are known to synthesize these novel compounds.

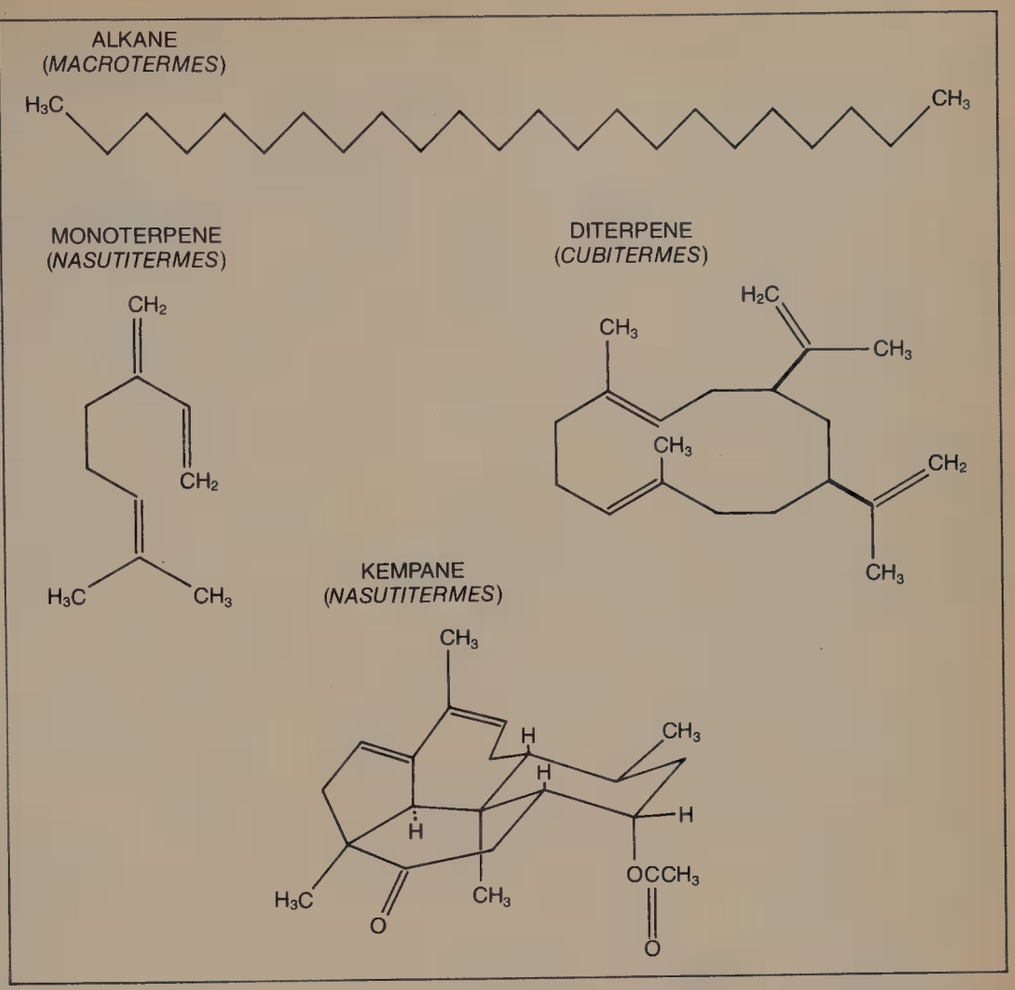
The third biting/injecting genus of higher termites is *Armitermes,* a soil-eating termite found in Central and South America. The soldiers of this genus have sharp-pointed mandibles resembling a pair of tongs and have a nozzlelike protrusion on their forehead. With the nozzle they apply droplets of an oily secretion to the puncture wounds they inflict on intruders. Thanks to diligent collecting in the rain forest of Guyana by one of my collaborators, Margaret S. Collins of Howard University, I have had access to soldiers of several *Armitermes* species for analysis of their secretions.

The *Armitermes* secretions consisted of modified fatty acids with from 22 to 36 carbon atoms to a molecular chain. One end of the chain bears a carboxylic acid group and the other end a hydroxyl group. Similar chains are found in the lanolin of wool, but what is unusual in these secretions is that the two ends of



SOLDIERS OF ANOTHER SPECIES (*Nasutitermes kempae,* found in Africa) cluster around the entrance to a foraging tube that has been opened to allow for photography. The snouts of four of the soldiers point in the general direction of the camera. This orientation is not accidental. Although termites are blind, the soldiers have oriented in response to air currents caused by the photographer's movements.

the chain are bound together in an ester linkage, forming a molecular loop. Similar loops with fewer carbon atoms, called macrolides, are found among the sex-attractant substances of mammals and substances that line the nests of solitary bees. As far as is known, however, macrolides with from 22 to 36 carbon atoms are unique to *Armitermes* and a related biting/injecting higher-termite genus, *Rhynchotermes*. Barbara L. Thorne of Harvard University and James F. A. Traniello of Boston University have found that when the secretion is applied to an intruder's puncture wound, it acts as a toxin rather than simply inhibiting the healing of cuticle.



TYPES OF MOLECULES found in termite secretions are represented by their backbones. Termites of the genus *Macrotermes* secrete alkanes: straight-chain hydrocarbons (*top*). Termites of the genera *Nasutitermes* and *Cubitermes* secrete more complex terpene hydrocarbons (*middle*). Even more complex terpenes are secreted by the most advanced *Nasutitermes*.

The second of the three main methods of chemical defense among termite soldiers, as established by Quennedey and Deligne, involves daubing an intruder with a toxic secretion of chemically reactive lipids derived from fatty acids. Soldier termites of the lower-termite family Rhinotermitidae are morphologically adapted for this mode of defense. Their labrum has been enlarged so that it forms a kind of bristly paintbrush the soldier presses against the intruder, spreading the pungent, oil-soluble contact poison over its cuticle. Unlike the biting/injecting higher termites, these species do not store their secretions in the head alone but have developed a large backup reservoir in their abdomen. In some soldiers the combined capacity of these reservoirs exceeds 35 percent of their dry weight.

Quennedey and his associates at Dijon first described the poison of an African genus, *Schedorhinotermes*, as a vinyl ketone in 1973. At about the same time Ján Vrkoč and his colleagues in the Czechoslovak Academy of Sciences reported that the poison of the pantropical genus *Prorhinotermes* was a nitroalkene. Since then my colleagues and I have confirmed and expanded both identifications in related species. We have also reported a third group of poisons, beta-ketoaldehydes, secreted by the soldiers of two New World Rhinotermitidae: the genera *Rhinotermes* and *Acorhinotermes*.

Although each of these contact poisons has a different molecular structure, all three have two key features in common. First, they are derived from fatty acid molecules with from 14 to 16 carbon atoms, and they all take the form of long carbon chains that are soluble in fat. Second, all three have a chemically reactive electrophilic (electron-seeking) group of atoms at one end of the chain. They might be likened to a poisoned arrow. The fat-soluble shaft of the "arrow" facilitates its passage though the intruder's waxy cuticle. Once inside the cuticle, the poisoned tip of the "arrow," the electrophilic group, does internal chemical damage. Such electrophilic groups are a common feature in natural defensive secretions. For example, they are found in the secretions of many fungi and higher plants. They can act as antibiotics, tumor inhibitors and insect repellents.

How do these termites avoid poisoning themselves with the toxins they secrete? They must have evolved biochemical adaptations that enable them to detoxify their own electrophiles; in the absence of some method of detoxification the termite soldiers could survive neither the synthesis nor the storage of their chemical weapons, and the termite workers could not survive their deployment. Until recently, however, no one had determined the exact biochemical basis of the termites' immunity. It was speculated that glutathione (GSH), a sulfur-containing tripeptide, and a group of detoxifying enzymes, the glutathione S-transferases, were involved. GSH S-transferases make electrophiles more water-soluble and less chemically reactive, thereby facilitating their excretion. Both vertebrates and invertebrates, and plants too, are known to rely on GSH S-transferase to rid their system of toxic substances. Moderate levels of GSH and GSH S-transferase are also found in termites.

One of my students, Stephen Spanton, and I discovered, however, that the workers of the Florida termite *Prorhinotermes simplex*, whose soldiers synthesize toxic nitroalkenes, and the workers of the African termite *Schedorhinotermes lamanianus*, whose soldiers synthesize toxic vinyl ketones, detoxify their soldiers' secretions in another way. They reduce the molecules' electron-poor double bonds, converting the nitroalkenes into nitroalkanes and the vinyl ketones into saturated ethyl ketones. The reduced products are only a tenth as toxic as the unreduced secretions.

We found that the initial detoxification was accomplished by an enzyme that specifically reduces the electrophilic alkene of that species and that it required NADPH, a reduced nucleotide cofactor, as a source of hydride. Among *Schedorhinotermes* workers 50 percent of the vinyl ketone was converted into ethyl ketone in less than an hour. Then all the ethyl ketone was catabolized (broken down) into acetate within two days. The *Prorhinotermes* workers enzymatically reduced their soldiers' toxin in an analogous way. When these workers were exposed to the toxin of the *Schedorhinotermes* soldiers, however, they could not tolerate doses as large as those

BREAKTHROUGH: ADD SOUND TO JET ENGINES TO MAKE AIRPLANES QUIETER.

Two airstreams exit from a jet engine in operation—a hot, fast central core and a cooler, slower outer layer. When they mix, they become chaotic. This turbulence causes an unpleasant low-frequency rumble.

We're finding new ways to control the turbulence by adding sound to smooth the flow. The smoother the flow, the quieter the airplane. The gentle symmetry you see here could mean a quieter ride in the future—quieter for the people aboard the plane, quieter for the people on the ground below.

We're making breakthroughs not only in aerospace but also in health care, information processing and energy.

We're McDonnell Douglas.

MCDONN
DOUGL

Artist's interpretation of turbulence control test.



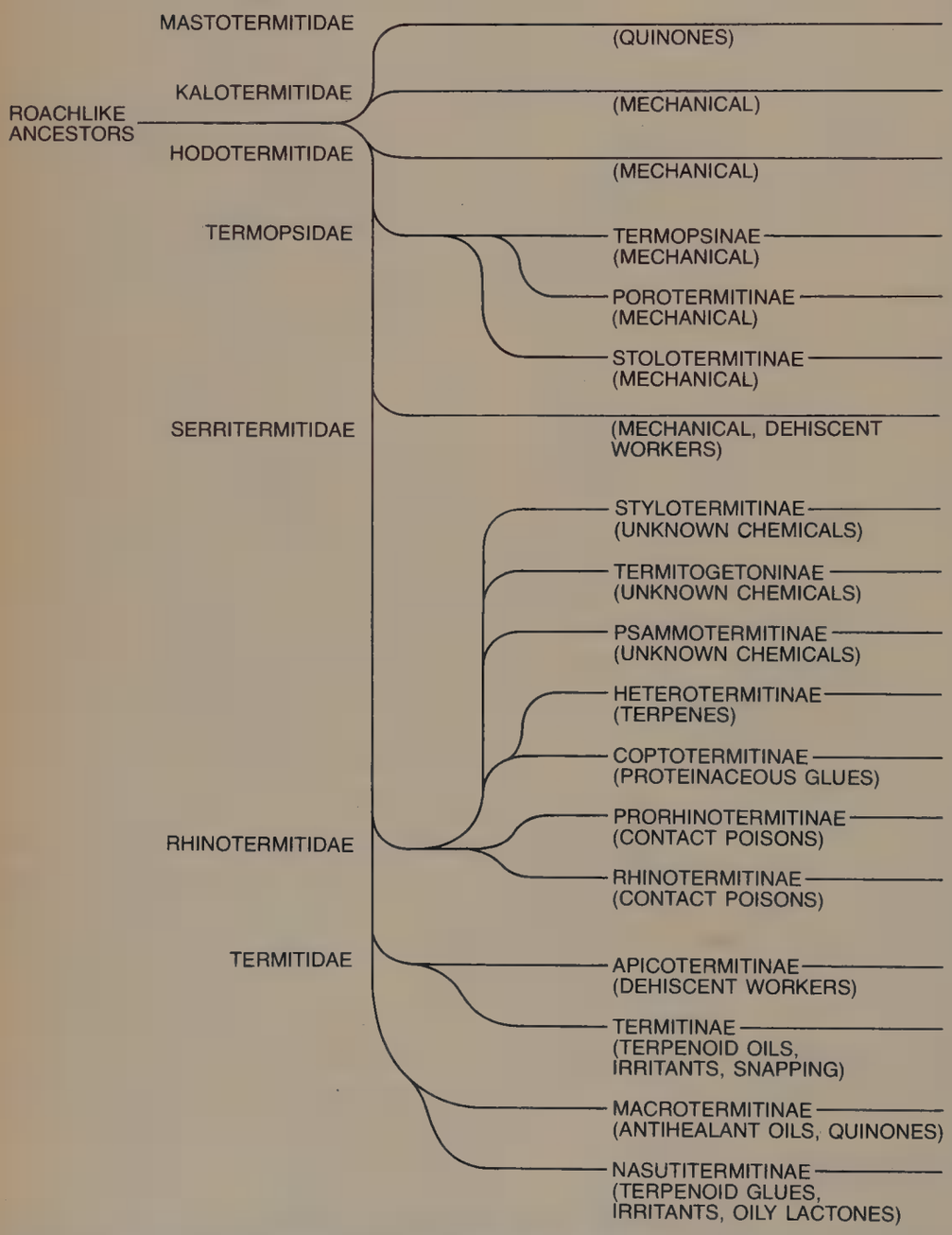
EL
AS

they survived when they were exposed to their own soldiers' toxin.

Except among these two termite species the reduction-catabolism pathway for detoxifying electrophilic poisons is apparently a rarity. This may be because termites in general are under evolutionary pressure to develop other detoxification pathways in order to conserve nitrogen for protein synthesis: their cellulose-rich foods are poor in nitrogen. For example, termites practice such nitrogen-conserving strategies as selective foraging and cannibalism.

Cathy J. Potrikus and John A. Breznak of Michigan State University have shown that the termites' own symbiotic bacteria can fix atmospheric nitrogen and recycle the nitrogenous compound uric acid; both activities contribute importantly to the termites' nitrogen supply. Barbara L. Bentley of the State University of New York at Stony Brook and I have shown that the *Nasutitermes* genera of Costa Rica fix enough molecular nitrogen to double their nitrogen supply in less than six months. The Floridian and African electrophile-resistant termites, however, appear to avoid nitrogen loss during detoxification by recycling, rather than excreting, both the nitrogen-rich amino acids used for detoxification and the energy stored in the carbon chain of the defense secretion.

The last main method of termite chemical defense, and in an evolutionary sense the most advanced of the three, calls for still another modification of the soldiers' anatomy. Instead of daubing an intruder with a toxin the soldiers of some 500 species in one abundant pantropical termite subfamily rely on a modified forehead with a nasus: a snoutlike tube. That subfamily is the Nasutitermitinae, named for the nasus. When a nasute nest is attacked, the alarmed soldiers charge from the interior to the point of intrusion and squirt an irritating gluelike secretion from their nasus to entangle the intruder. The workers join the action, biting at the intruder with their sharp mandibles and trying to clamp onto its legs. Ants, spiders and other insectivorous animals, including anteaters, generally avoid the malodorous, distasteful and potentially lethal barrage of the nasute soldiers.

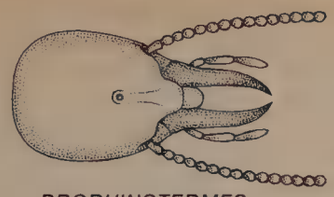


SEVEN FAMILIES OF TERMITES arose (*left*) from a roachlike ancestral stock. The six families known collectively as the lower termites depend mainly on mechanical defenses: the powerful jaws of their soldiers. In this group, however, the Rhinotermitidae include seven subfamilies, six of them possessing soldiers that use chemical defenses either with biting or without. The seventh termite family, the Termitidae (*bottom*), includes the four subfamilies of higher termites. One subfamily, the Apicotermitinae, has no soldier caste. Colony defense is undertaken by workers that are dehiscent: they explode and shower the intruder with gut contents. The Nasutitermitinae are found in the Tropics of both the Old World and the New. They evidently evolved some 70 million years ago. Defenses of each group are given in parentheses.

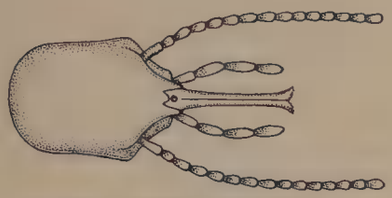
It has been known for some time that the viscid defensive secretion of the nasutes was analogous to pine sap: a mixture of monoterpene hydrocarbons (the solvent) and other isoprenoids of higher molecular weight (the resin). The termite secretion differs from pine sap, however, in containing none of the usual resin acids. In 1974, when my colleagues and I began to investigate this nasute glue, we had many questions. What exactly were the heavy compounds? Were they merely sticky or were they toxic as well? Did the soldiers actively biosynthesize the substances or were they acquired from food? The nasute termites make up the most numerous and diverse termite subfamily and are worldwide in their distribution; how did the composition of the glue vary from species to species and from place to place? Now, nine years later, we are able to answer some of these questions.

Our first success came with the isolation of several glue components in purified form and the elucidation of their structure. This required the combination of my work in Africa with the efforts of Jon C. Clardy and his colleagues at Iowa State University and Koji Nakanishi and his colleagues at Columbia University. Because thousands of nasute soldiers of the African grass-harvesting genus *Trinervitermes* were the source of the research material, the first glue components were named trinervitanes. They are diterpene molecules consisting of three fused rings with their skeleton of carbon atoms in a dome-shaped arrangement. The carbon skeletons do not resemble any other known natural molecular structure.

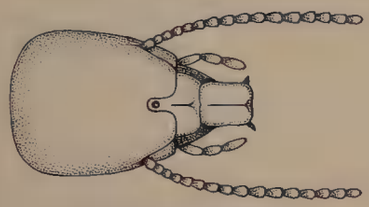
Since then we and others have isolated one bicyclic (two-ring) and three tetracyclic (four-ring) diterpenes, known respectively as secotrinervitane, kempane, rippertane and longipane. Together with some 60 derivative substances they are secreted by the soldiers of other nasutetermite genera. All exhibit the same



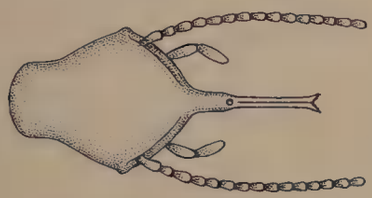
PRORHINOTERMES



RHINOTERMES (MINOR)



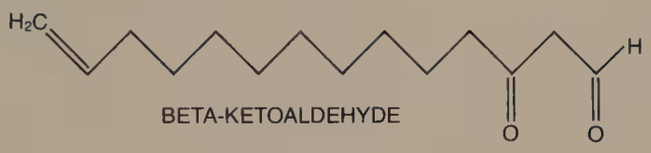
SCHEDORHINOTERMES



ACORHINOTERMES



NITROALKENE



BETA-KETOALDEHYDE



VINYL KETONES



BETA-KETOALDEHYDE

MODIFICATIONS OF THE HEAD in the soldiers of four termite genera in the family Rhinotermitidae are shown from the top at the left. The molecular structure of the defensive secretions that each soldier applies to an invader of its colony is at the right. The long-chain hydrocarbons secreted by the four genera range from a simple nitroalkene to a vinyl ketone and a more reactive beta-ketoaldehyde.



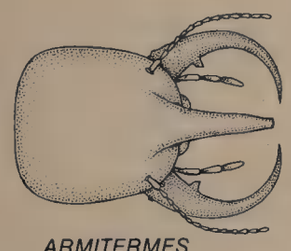
MACROTERMES



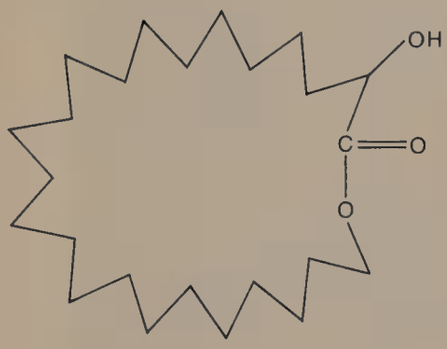
QUINONES



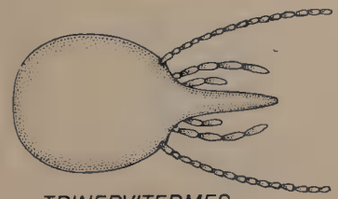
ALKENES



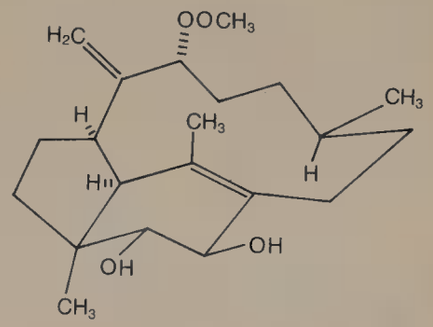
ARMITERMES



MACROCYCLIC LACTONE



TRINERVITERMES



TRINERVITANE

SIMILAR MODIFICATIONS in the heads of soldiers have arisen in three higher-termite genera. At the bottom is the modification typical of the nasute genus *Trinervitermes*. The *Macrotermes* secretions can be simple quinones, alkanes or alkenes. The nasute genus *Armitermes* secretes molecular loops called macrolides. The *Trinervitermes* secretion is a mix of diterpenes unknown elsewhere in nature.



SQUARED HEAD is the adaptation of the *Cryptotermes* soldier. When these soldiers are alarmed by an intruder, they plug the entrance to the colony with their heads, presenting an array of mandibles to discourage other intruders. These termites have no chemical defenses.



SABERLIKE MANDIBLES of a *Cubitermes* soldier will gash the cuticle of an attacking ant while a secretion from the soldier's frontal gland is spread on the wound. The secretion seems to inhibit the intruder's blood coagulation and cuticle repair, and it will soon bleed to death.

gross structural features. Each molecule has the same biosynthetic progenitor: a monocyclic diterpene, cembrene-A. Each is dome-shaped, with oxygen-containing water-attracting groups protruding from the convex surface and a water-repelling region on the concave surface. Defense against intruders is thus achieved with a substance rather like pine sap: a viscous solution of mixed diterpenes in association with monoterpene solvents. The mixture has superior wetting abilities when it is applied to the normally water-repellent cuticle of insects and other arthropods.

We have also found that the secretions vary among individual soldiers in a single colony, among soldiers in related populations of the same species, among geographically isolated populations of the same species and among species of the same genus. This kind of chemical variation, which is commoner in the plant kingdom than it is in the animal one, can be useful in taxonomic studies.

Exactly what does the nasute glue do? Thomas Eisner and his colleagues at Cornell have shown in a series of elegant laboratory experiments that it functions as an entangling agent, as an irritant (a property that promotes its spread when the sprayed intruder grooms itself) and as a topical poison. Whether the soldiers synthesized the glue or obtained it from food remained a question my colleagues and I decided to pursue with radioactive-tracer experiments. One cannot feed a labeled precursor to termite soldiers because the only nourishment they will accept is the fluid that workers regurgitate for them. We bypassed the problem by resorting to a micropipette to inject radioactively labeled precursor molecules (sodium acetate or sodium mevalonate) directly into the soldiers' abdomen. Because the termite abdominal wall is very flexible, we were able to inject about half a microliter into each subject, roughly doubling the size of its abdomen.

The soldiers survived for some hours after the injection. By then each precursor had been incorporated into the monoterpenes and diterpenes stored in each soldier's head. When we injected workers (which do not secrete diterpenes) with the same labeled precursors, the precursors were not found in their terpenoid compounds. After careful purification of each component we concluded that the nasute soldiers were indeed synthesizing their defensive secretions de novo. Their ability to do so sets them apart from all other insects (except perhaps for certain scale insects). None of the rest can synthesize any terpenes larger than a 15-carbon-atom sesquiterpene.

Another experiment demonstrated that the diterpenes secreted by a nasute soldier were not affected by the soldier's

diet. The method involved raising two chemically distinct populations of a single *Trinervitermes* species, beginning with separate pairs of winged male and female reproductives after their nuptial flight. Both pairs and the colonies they gave rise to were given identical foodstuffs: dried grass from one of the collection sites. In spite of their common diet the soldiers that reached maturity six months later in each new colony produced diterpenes chemically identical with those of their parent colonies.

The chemical analyses of the nasute soldiers' secretions have illuminated a particular aspect of termite evolution. It has been customary to consider the pantropical nasute subfamily, the Nasutitermitinae, a leading example of parallel evolution. This is to say that the nasute genera of Africa, Asia, Australia and South America arose from ancestral termites whose soldiers were armed with mandibles alone. In the course of thousands of generations the primitive soldiers (now found only in South America) underwent a regression of their mandibles and a modification of their forehead into the bazooka-like nasus. It seems to be asking too much of a parallel evolutionary process, however, to go on to say that the capacity to synthesize defensive diterpenes, a feat of chemistry unique to the advanced nasute termite soldiers, also evolved independently. I suggest that instead a common ancestral stock of diterpene manufacturers existed in West Gondwana before plate tectonics divided that land mass into a proto-Africa and a proto–South America in Cretaceous times.

In support of this contention one can point to the fact that many existing "primitive" nasute genera can be fitted into an ancestral scheme in which the mandibles did shrink and the nasus did elongate. Chemical examination of the secretions of the intermediate genera, however, fails to show any of the four unique diterpenes that characterize the secretions of the advanced nasutes. The South American nasutes with large mandibles cannot squirt their secretions; they can only ooze them out in droplets, and what is oozed consists of fatty-acid-derived macrolides and mono- and sesquiterpenes. The more "advanced" South American nasutes, on the other hand, do not have functional mandibles but can squirt diterpene solutions as their African, Asian and Australian relatives do. The identical three-dimensional chemical structure of all the nasute diterpenes worldwide supports the hypothesis of a common evolutionary origin. Thus it seems that what appeared to be an excellent example of parallel evolution must be written off the books. It also seems likely that further study of the defensive chemistry of the termites in general will yield still further surprises.



SNOUTED TERMITE SOLDIER of the genus *Armitermes* is a similar chemical defender. After piercing an attacker's cuticle with its hooked mandibles it uses its snout to apply droplets of a greasy secretion to the wounded area. The mixture is toxic when it enters the wound.



LONG UPPER LIP of a soldier in the genus *Rhinotermes* is used to daub the cuticle of an intruder with a toxic ketoaldehyde stored in its head and abdomen. The fat-soluble poison then penetrates the intruder's waxy cuticle and fatally interferes with metabolic processes.

Rational Collective Choice

Axiomatic analysis of voting systems has probed the compatibility of several desirable properties of an ideal method. Compromises among rationality, decisiveness and equality seem unavoidable

by Douglas H. Blair and Robert A. Pollak

Can a system of voting be devised that is at the same time rational, decisive and egalitarian? Studies of this question by philosophers, political scientists and economists (including the two of us) suggest that the answer is no. These characteristics of an ideal system are in fact incompatible. A method of voting may avoid arbitrariness, deadlock or inequality of power, but it cannot escape all three. The continuing analysis of this dilemma has led to a deeper understanding of existing voting systems and may lead in time to the discovery of better ones.

The axiomatic analysis of rational voting procedures was initiated some 33 years ago by the economist Kenneth J. Arrow of Stanford University. He advanced five intuitively appealing axioms that any procedure for combining or aggregating the preferences of individuals into collective judgments should satisfy, and he proved that the only procedures obeying all of them concentrate all power in the hands of a single individual. No nondictatorial method satisfying all Arrow's axioms can be found, not for want of ingenuity but because none exists. In part for this work Arrow shared the Nobel prize in economics in 1972.

Over the past 15 years investigators have reexamined Arrow's axioms in an effort to circumvent his "impossibility theorem" by relaxing his requirements. The problem has attracted widespread interest because it is closely linked with central questions in philosophy, political science and economics. Philosophers face it, for example, in analyzing the practical implications of utilitarianism, the ethical doctrine holding that the rightness of actions depends on their consequences for people's happiness and hence requiring a method for aggregating the preferences of individuals. Political scientists encounter it in designing or evaluating rules of voting for committees or legislatures. Economists confront it in analyzing rationing and other nonmarket methods of allocating resources. This task is an important one in normative economics, because in determining the appropriate scope for intervention by the government in the operations of a free-market economy it is crucial to understand the potential performance of the alternatives to laissez faire.

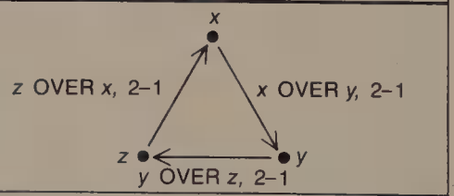
Majority rule deserves first consideration among procedures for aggregating individuals' preferences; its virtues include simplicity, equality and the weight of tradition. Majority rule is essentially a procedure for ranking pairs of candidates or alternatives. When more than two alternatives must be ranked, however, majority rule encounters a difficulty the Marquis de Condorcet recognized nearly 200 years ago.

The difficulty pointed out by Condorcet is now known as the "paradox of voting." Suppose a committee consisting of Tom, Dick and Harry must rank three candidates, x, y and z. Tom's preference ranking of the candidates is x, y, z, Dick's is y, z, x and Harry's is z, x, y. Majority voting between pairs of candidates yields a cycle: x defeats y, y defeats z and z defeats x, all by two votes to one. This voting cycle is the simplest example of Condorcet's paradox of voting.

Political scientists have identified many historical cases of voting cycles. For example, William H. Riker of the University of Rochester argues that the adoption of the 17th Amendment, providing for the direct election of U.S. senators, was delayed for 10 years by parliamentary maneuvers that depended on voting cycles involving the status quo (the appointment of senators by the state legislature) and two versions of the amendment.

When more than two alternatives are feasible, some new principle is needed for generating choices from pairwise rankings. The preference configurations that induce the paradox of voting create difficulties for each natural approach. The simplest method chooses an alternative that is undefeated by any other. In a paradox-of-voting situation, however, no such alternative exists, because each alternative or candidate loses to another.

A second method for proceeding from pairwise rankings to choices is to specify an agenda, listing the order in which pairs of alternatives will be taken up. For example, the agenda might call for an initial vote on x v. y, followed by a second stage in which the winner is matched against z. Under this agenda our three-member committee would first vote for x over y and at the second stage z would defeat x. It is easy to verify that under each of the three possible agendas in this situation the alternative taken up last emerges as the victor: the agenda determines the result. Voting cycles therefore present substantive difficulties as well as aesthetic ones. When a cycle occurs, the choice of an ultimate winner is at best arbitrary (if the agenda is selected randomly) and at worst deter-

COMMITTEE MEMBERS	PREFERENCE AMONG ALTERNATIVES	OUTCOME OF MAJORITY VOTING
TOM	x, y, z	
DICK	y, z, x	
HARRY	z, x, y	

PARADOX OF VOTING can arise under majority rule when voters with conflicting preferences must choose among more than two alternatives. The paradox is depicted for a three-member committee considering three candidates or alternatives, x, y and z. The outcome is cyclic: x defeats y, y defeats z and z defeats x, all by two votes to one. Such problems led Kenneth J. Arrow of Stanford University to propose five axioms that a voting procedure should satisfy.

mined by the machinations of the agenda setter.

Further opportunities for strategic maneuvering arise if a voter can alter the agenda by introducing new alternatives. Suppose (with the same committee) z represents the status quo and y is an alternative embodied in a motion that has been introduced. With only these two alternatives available y will defeat z, and Harry (who prefers z to y) will be disappointed. If he can introduce an amendment x to the motion y, however, x will defeat y on the initial vote, and at the second stage x will lose to z. Harry thus will obtain the enactment of his favored alternative.

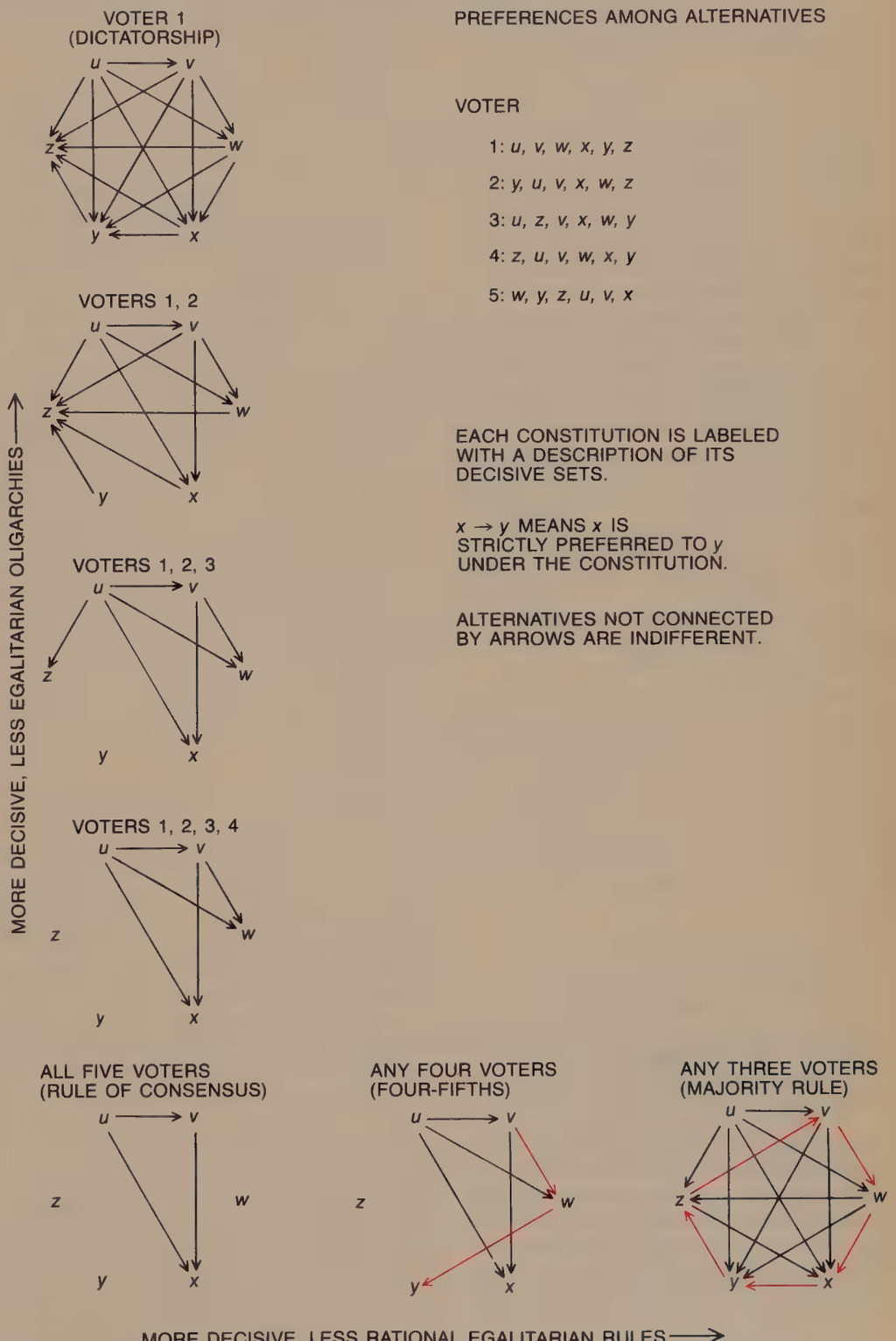
Even if new alternatives cannot be introduced and the agenda cannot be manipulated, opportunities may still exist for voters to profit by misrepresenting their preferences. Consider again the agenda in which z is taken up last. If each member of the committee votes his true preference on each ballot, the winning alternative, z, is the least desirable one for Tom. Suppose, however, Tom votes for y instead of x on the initial ballot; then y prevails, going on to defeat z in the second stage. By this stratagem he has blocked the choice of the alternative he liked least.

Cyclic collective preferences present problems of arbitrary outcomes and strategic behavior. These difficulties arise whether the preferences are generated by majority rule, as in our example, or by some other voting procedure. Arrow was therefore led to ask: Do inconsistent collective preferences arise only under majority rule and closely related methods or are they inherent in all voting systems? To answer the question he might have assembled a list of voting procedures and for each procedure checked whether any configurations of individuals' rankings gave rise to cycles or to collective preferences with some other unacceptable feature. The difficulty is that he would have had to consider an immense number of aggregation procedures, differing widely in the roles they assign to particular voters and in the criteria they employ in ranking particular pairs of alternatives.

Of necessity Arrow chose an axiomatic approach instead. He formulated the problem as the choice of a constitution, that is, a rule assigning a collective ranking of the alternatives to each configuration of individuals' rankings. A constitution specifies whether each alternative stands as preferred, inferior or indifferent to every other one. (Two alternatives are indifferent if society regards them as being equally attractive.) Arrow narrowed the field of possible constitutions by imposing five requirements that (he argued) are necessary properties of any ethically acceptable method of aggregation. He then characterized the class of constitutions that satisfy all five properties.

The first of Arrow's axioms, Universal Scope, requires that a constitution be capable of aggregating every possible configuration of voters' preferences. Since one cannot predict all the patterns of conflict that will arise over the life of a voting rule, Arrow argued, a society should not adopt a constitution that will break down when certain configurations of voters' preferences arise. He contended that the society should instead insist on a constitution sufficiently general to resolve all possible controversies.

Arrow's second axiom, Unanimity, governs the operation of a constitution when there is no disagreement among

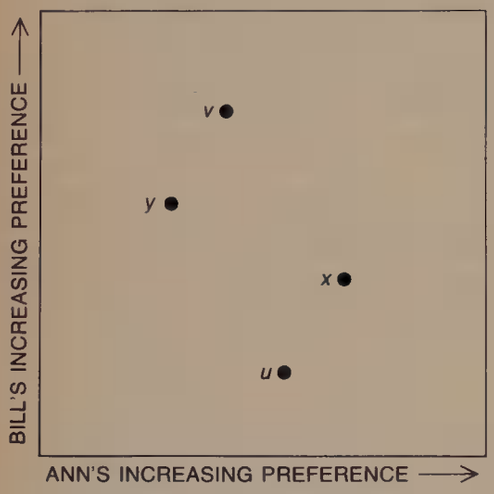


COLLECTIVE-PREFERENCE RANKINGS corresponding to a particular set of preferences held by five voters are shown for seven constitutions, or voting rules. The rankings typify the tradeoffs faced by the designer of voting rules among rationality, inequality of power and decisiveness (the infrequency of collective indifference). Majority rule gives voters symmetric roles and is highly decisive, but for the preference rankings of these five voters a five-way cycle (*color*) and several shorter cycles arise. The rule of four-fifths is less decisive but yields no cycle here. It fails, however, to be *P*-transitive (consistent in ranking strict preferences) in that (to employ a shorthand notation) *vPw* (alternative *v* is strictly preferred to alternative *w*) and *wPy* but *vIy* (*v* is indifferent to *y*, meaning that society views them as being equally attractive). Oligarchies are always *P*-transitive but are less decisive as they get larger and more egalitarian.

voters. It specifies that for preference configurations in which every individual prefers x to y the collective ranking must put x above y. If one accepts the view that a society's ranking should reflect its members' preferences, it is difficult to quarrel with the Unanimity condition, which resolves what surely are the easiest problems of preference aggregation.

Arrow's third axiom, Pairwise Determination, requires that society's ranking of any pair of alternatives depend only on individuals' rankings of those two alternatives. No matter how the preferences of individuals for other alternatives may change, as long as each individual's ordering of x and y remains invariant the collective ranking of x and y does also. This condition implies, for example, that the collective ranking of Ronald Reagan and Jimmy Carter is independent of how individuals rank Edward Kennedy with respect to those two or to Walter Mondale.

A constitution satisfying Pairwise Determination limits the information about individuals' rankings required to determine the collective ranking of a pair of alternatives. In particular, information about the preferences of individuals for unavailable options is irrelevant to the collective ranking of the available ones; this is an advantage when it is difficult or costly to elicit individuals' preference rankings. Without the condition of Pairwise Determination the constitution must specify what other alternatives are relevant to the determination of the collective ranking of x and y and how the preferences of individuals for those alternatives affect the collective ranking of x and y.



P-TRANSITIVE CONSTITUTIONS are voting rules in which if xPy and yPz, then xPz. All such constitutions satisfying Arrow's other axioms are neutral, that is, they rank pairs of alternatives by the same criteria. Neutrality means that if a particular configuration of voters' rankings of u and v implies that u is collectively preferred to v, the same configuration of rankings of x and y implies that x is collectively preferred to y. A proof with two voters is depicted. Suppose that for the alternatives u and v Ann prevails over Bill under the constitution when she prefers u to v. On x and y they differ. For this configuration xPu by Arrow's Unanimity axiom, uPv by the assumption that Ann prevails for u against v and vPy by Unanimity. Under P-transitivity it follows that xPy for this particular preference configuration. Arrow's Pairwise Determination condition, however, allows the more general conclusion that Ann prevails for x against y, regardless of the positions of u and v in her ranking or in Bill's. (Pairwise Determination states that the collective ranking of any pair of alternatives depends only on the preferences of voters regarding those two choices.)

One common procedure, rank-order voting, violates Pairwise Determination. (It is the system normally employed by newspaper wire services to determine the ranking of college athletic teams.) When there are three alternatives, this constitution assigns each individual's first choice three points, his second two points and his third one point; the collective ranking is then found by summing the scores for each alternative and ranking them according to their total scores. In the three-member committee we have described each candidate receives a score of six, and so the committee is indifferent among the three candidates.

Suppose, however, Tom's preference ranking changes from x, y, z to x, z, y. Although no voter has changed his ranking of x and y, rank-order voting now yields a collective ranking of x over y, since x still receives a score of six but y now gets five. Hence under rank-order voting the collective ordering of x and y depends not only on how individuals rank them but also on the relative positions of other alternatives such as z.

Arrow's fourth and fifth axioms are best discussed using a shorthand notation. P denotes a strict collective preference (analogous to the relation "greater than" between a pair of real numbers), I denotes collective indifference (analogous to equality) and R denotes a weak collective-preference relation (analogous to "greater than or equal to"). Thus the expression xRy stands for "x is collectively at least as good as y," that is, either xPy or xIy.

Arrow's fourth axiom is Completeness: for every pair of alternatives x and y it must be true that xRy or yRx (or both, in which case x and y are indifferent). This axiom compels the aggregation procedure to rank every pair of alternatives. As long as the constitution has the option of declaring any pair of alternatives indifferent, Completeness seems to be a relatively innocuous requirement.

The fifth axiom, R-transitivity, requires that a weak collective preference be transitive: formally, if xRy and yRz, then xRz. Transitive relations involving pairs of real numbers include "greater than" ($>$), "equal to" ($=$) and "greater than or equal to" ($\geq$). Hence if a number x is greater than y and y is greater than z, x must be greater than z. In economic analysis Completeness and R-transitivity are conventional assumptions, and individuals whose preferences obey these axioms are said to be "rational." By extension Arrow employed the term "collective rationality" to describe constitutions satisfying both Completeness and R-transitivity. He imposed R-transitivity to ensure that the chosen alternative would be independent of the agenda or path by which it is reached.

Having defined and defended this set of five desirable properties, Arrow proved that the only constitutions satisfying all of them share a simple and startling defect: each constitution is dictatorial. A dictator is a person with the power to impose on the society his strict preference over any pair of alternatives. Arrow stated his theorem in a slightly different way. He added a sixth axiom, nondictatorship, and proved that no constitution exists that obeys all six axioms. For this reason Arrow's result is often described as an "impossibility theorem."

Thus the designer of voting procedures for legislatures, committees and clubs who accepts these conditions as necessary properties of constitutions is simply out of luck. Arrow's apparently modest requirements have powerful and unpalatable implications. As his impossibility theorem demonstrates, the five axioms are highly restrictive; although they are attractive singly, they are pernicious in combination. Voting theorists have devoted much effort to reexamining the axioms, seeking a way around Arrow's unhappy conclusion.

A plausible argument can be made that Universal Scope is too ambitious a requirement. Not every logically possible configuration of preference rankings is equally likely. Since some configurations may be extremely unlikely, requiring a constitution to aggregate consistently every logically possible configuration into a collective ranking seems unnecessarily strong.

The commonest strategy in relaxing this requirement has been to focus on a particular procedure, usually majority rule, and to look for restrictions that rule out preference configurations implying intransitive collective preferences. For example, if only preference configurations in which there is no disagreement among individuals could arise, the Unanimity axiom would determine collective preferences and the problem of intransitivity could not arise. The best-known nontrivial restriction is single-peaked preferences, discovered in the 1940's by the British economist Duncan Black.

Single-peakedness arises when all individuals evaluate alternatives according to some single criterion and, in any pairwise choice, each individual votes

for the alternative closer to his own most preferred position. For example, each voter might rank candidates according to how close they were to his own position on the political spectrum from liberal to conservative. Hence if x is more liberal than y and y is more liberal than z, a society with single-peaked preferences that contained liberals (x, y, z), conservatives (z, y, x), and moderates (y, x, z or y, z, x) could not contain individuals for whom the middle alternative is ranked below both extremes (x, z, y and z, x, y). If single-peakedness could be expected to hold in practice, the case for majority rule would be compelling. Usually, however, people rank alternatives by multiple criteria and so single-peakedness will fail.

More generally, the strategy of imposing restrictions on preference configurations can be fruitful only if the restrictions are plausible in terms of a theory of preference formation or preference structure. Social scientists, however, have not succeeded in formally modeling either the role of socialization in the development of tastes and values or the degree of similarity of preferences needed for social stability. Therefore, notwithstanding a great deal of effort, no characterization of possible patterns of rankings has yet been formulated that is broad enough to encompass voters' actual preferences and at the same time narrow enough to evade the dictatorship conclusion.

The possibility of abandoning the Unanimity axiom has generated little enthusiasm. On reconsideration, Unanimity still seems to be a mild requirement to impose on mechanisms for aggregating individuals' preferences into collective rankings. Furthermore, Robert Wilson of Stanford University has shown that the only additional constitutions satisfying Arrow's other axioms but violating Unanimity are even less appealing than dictatorships. In particular there are two new possibilities. The first possibility is universal indifference, the rule that makes every pair of alternatives perpetually indifferent regardless of how individuals rank them. The second is inverse dictatorship, a rule under which some particular individual's preference ranking is reversed to form the collective ranking. This method would serve an orderly society only if a voter with infallibly bad judgment could be found.

Pairwise Determination drew heavy fire in the first decade after Arrow published his work, but criticism of this axiom has subsided. Arrow's original defense of the condition was that constitutions satisfying it can be implemented without the burden of collecting large amounts of preference information. To rank the alternatives x and y it is never necessary to ascertain the position of z in individuals' rankings. Constitutions that violate this axiom are generally cumbersome, at least when there are many alternatives, because much preference information must be obtained to rank even a small number of feasible alternatives. Moreover, since the collective ranking of x and y under constitutions violating this axiom is sensitive to individuals' rankings of third alternatives, it is often possible for voters to manipulate the outcome on x and y by misrepresenting their preferences about other alternatives.

Arrow's least defensible requirement is probably R-transitivity. To avoid Arrow's conclusion theorists of voting have examined the consequences of several less restrictive conditions. It is not difficult to show that R-transitivity is equivalent to the conjunction of two weaker conditions, P-transitivity (the transitivity of the strict collective-preference relation P) and I-transitivity (the transitivity of collective indifference). Therefore a straightforward way to weaken Arrow's rationality requirement is to retain one of these conditions while abandoning the other.

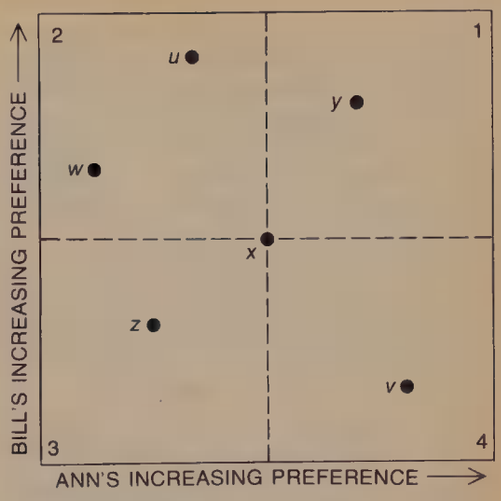
I-transitivity is particularly vulnerable to criticism, since studies by psychologists have shown that individuals in experimental situations often exhibit intransitive indifference. For example, an individual who expresses indifference between x and y and between y and z will often prefer x to z. Thus the analogy between individual preference and collective preference yields little support for requiring collective rankings to be I-transitive.

The economist and philosopher Amartya K. Sen of the University of Oxford showed that abandoning I-transitivity while retaining P-transitivity provides an escape from Arrow's dictatorship result. He offered an example of a nondictatorial constitution satisfying Arrow's first four axioms and P-transitivity. His procedure, which might be called the rule of consensus, yields xPy if and only if every individual ranks x as being at least as good as y and at least one individual strictly prefers x to y. Therefore when two individuals disagree over x and y, the result is xIy. Each individual has a veto that enables him to block a strict collective preference opposite to his own. To see that collective indifference need not be transitive consider a committee with two members, one with a preference y, x, z and the other with a preference x, z, y; the rule of consensus gives xIy and yIz but xPz.

The phenomenon of intransitive individual indifference may reflect the inability of individuals to distinguish among alternatives that are close together. For example, in judging political conservatism candidates x and y may seem equally conservative to an individual because their policy positions are indistinguishably close, and the same may seem true of y and z. Yet candidates x and z may be far enough apart for the individual to perceive that x is more conservative than z and hence to prefer one to the other.

The psychologist R. Duncan Luce of Harvard University has proposed a notion of consistency called the semiorder to model situations entailing such thresholds of perception. Semiordered preferences exhibit P-transitivity but allow intransitive indifference. If collective choice is seen as a process of aggregating the policy judgments of individuals to form collective policy judgments, the notion of imperfect discrimination can be applied to collective preferences as well as to individual ones.

The semiorder, however, is a stronger rationality requirement than P-transitivity alone, a fact that has important consequences for preference-aggregation procedures. As we have demon-

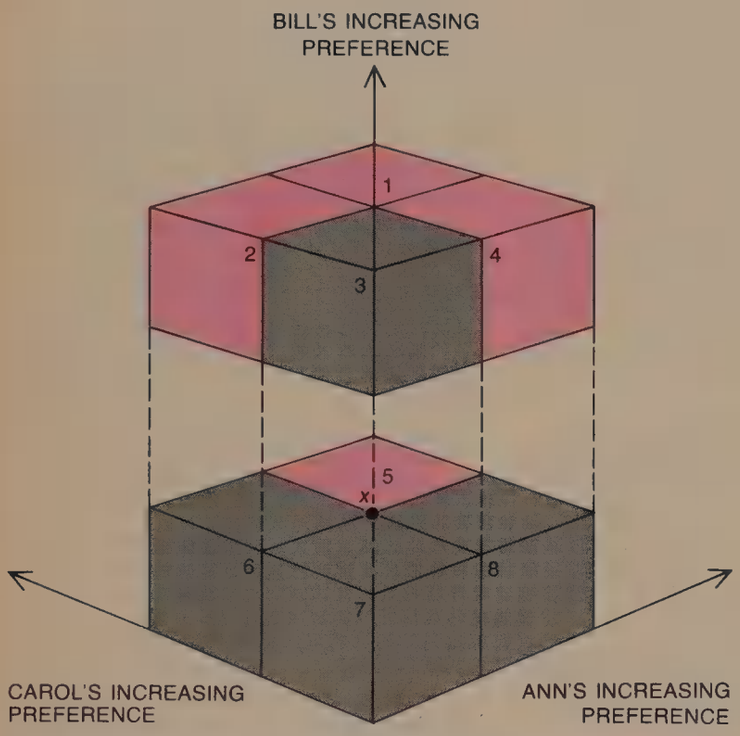


DICTATORSHIP THEOREM of Arrow is proved for two voters. One possible configuration of preference rankings for Ann and Bill is depicted; the quadrant in which each alternative appears depends on how the two rank it with respect to the alternative x. Quadrant *1* contains alternatives (such as y) both prefer to x, *2* the ones Bill prefers to x but Ann does not, *3* the ones neither prefers to x and *4* those Ann prefers but Bill does not. The neutrality property of transitive constitutions implies that every alternative in a quadrant must be ranked collectively in the same way with respect to x. It also implies that the argument does not depend on the particular alternative defining the quadrants. Quadrant *2* cannot be indifferent to x for then both uIx and xIw. Thus by I-transitivity uIw, contradicting Unanimity, since both voters prefer u to w. Similarly, quadrant *4* cannot be indifferent. Quadrants *2* and *4* cannot both be preferred to or inferior to x, since Neutrality implies that if uPx, then xPz, because the voters' rankings of u and x and of x and z are the same. Since Arrow's Unanimity axiom implies that quadrant *1* is preferred to x and *3* is indifferent, Arrow's five axioms are consistent with only two constitutions: if *4* is preferred to x and *2* is inferior, Ann is a dictator, whereas if *2* is preferred and *4* is inferior, Bill is a dictator. The proof is due to Charles Blackorby, David Donaldson and John Weymark.

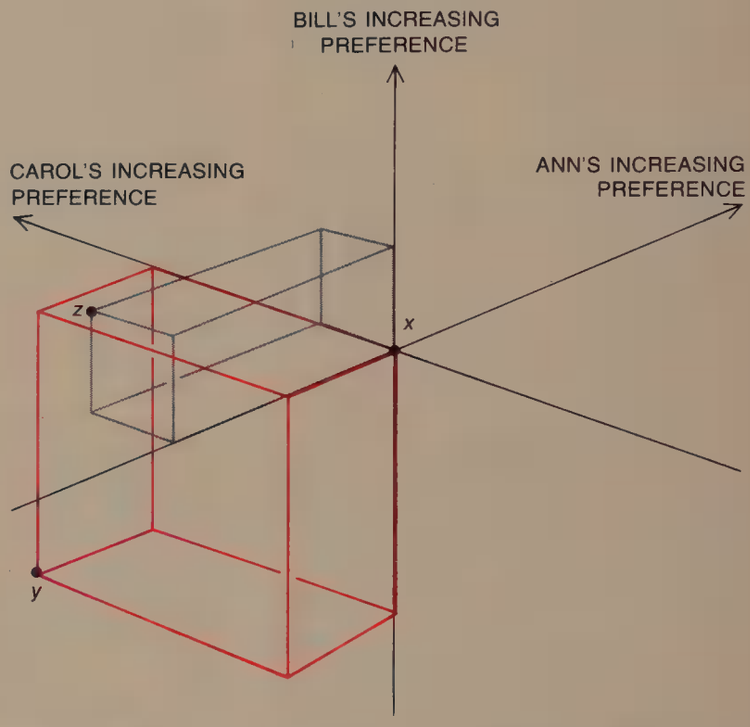
strated elsewhere, requiring that a constitution yield semiordered collective preferences and satisfy Arrow's other axioms still implies dictatorship. Thus the perception-threshold justification for weakening *R*-transitivity, although it is appealing and plausible, does not avoid Arrow's dismal conclusion.

Arrow's principal justification for *R*-transitivity was that the choice from some set of alternatives should be independent of the agenda or path by which the choice is made. Remarkably, the desire for path independence leads directly to an argument for *P*-transitivity alone as the appropriate rationality requirement for collective-preference rankings. Charles R. Plott of the California Institute of Technology has proposed a formal definition of path independence and has shown that all constitutions satisfying this condition are *P*-transitive. Furthermore, any *P*-transitive constitution satisfies path independence. Therefore, although Arrow's original collective rationality condition guarantees path independence, a less restrictive condition would accomplish the same objective.

Sen's rule of consensus shows that nondictatorial *P*-transitive constitutions exist. As Sen recognized, however, the rule of consensus is not an appealing solution to Arrow's problem because it is so often indecisive. Whenever any two individuals have opposing strict preference rankings of a pair of alternatives—surely a ubiquitous form of conflicting



NEUTRAL CONSTITUTIONS for three voters can be represented with eight-octant diagrams of voters' preferences analogous to the four-quadrant diagram for two voters. By neutrality, which is assumed for acyclic, or noncycling, constitutions and is a necessary property of *P*-transitive ones, all the alternatives in an octant must bear the same relation (preferred, indifferent or inferior) to x in the collective ranking. If one octant, such as *2*, is collectively preferred to x, its opposite (*8*) must be inferior, and if an octant is collectively indifferent to x, its opposite is also indifferent. Since the Unanimity axiom compels a constitution to prefer octant *1* to x and x to octant *7*, all possible neutral constitutions are fully described by specifying the collective preferences (with respect to x) for octants *2*, *3* and *4*. Since each of the three octants can be preferred, indifferent or inferior to x, there are 27 possible three-person neutral constitutions. Under *P*-transitive constitutions satisfying Arrow's other axioms it can be shown that additional support for x with respect to y cannot worsen x's position with respect to y in the collective ranking. For example, suppose Ann changes her ranking so that z moves from octant *3* to octant *4*, that is, instead of preferring x to z she now prefers z to x. Then the collective ranking of *4* with respect to x must be at least as favorable as *3;* if *3* is indifferent, *4* must be either indifferent or preferred. Some constitutions that are not *P*-transitive fail to satisfy this property of positive association of the collective ranking with individuals' rankings. Since such a constitution is unattractive, positive association can be made an additional requirement for constitutions that violate *P*-transitivity. Because the movement of an alternative from *3* to *2* or *4* reflects an improvement in some voter's evaluation of the alternative with respect to x, positive association requires that if *3* is preferred to x, *2* and *4* are too. If *3* is indifferent, *2* and *4* must be either preferred or indifferent. If *2* is inferior, *8* must be preferred; hence *4* must be preferred by positive association. Such arguments reduce the number of three-person constitutions satisfying neutrality, positive association and Arrow's first four axioms from 27 to 11. Majority rule is the one depicted here, with preferred octants in color and inferior ones in gray. No octant is indifferent under a majority-rule constitution.



***P*-TRANSITIVE AND ACYCLIC CONSTITUTIONS** can be characterized by means of the configuration of individuals' preferences shown here. There are 11 three-voter constitutions that satisfy neutrality, positive association and Arrow's first four axioms; they are represented in the illustration at the left and the one on the opposite page. Every acyclic constitution must have an individual with veto power. In other words, no octant can be collectively preferred to x if the vetoer prefers x to that octant. Each of the 11 constitutions except majority rule has at least one vetoer. Showing that majority rule can cycle proves the result. Under majority rule octants *1*, *2*, *4* and *5* are collectively preferred to x, whereas *3*, *6*, *7* and *8* are inferior. For the configuration of individuals' preferences indicated here majority rule cycles: xPy (because y is in octant *6*), yPz (since y bears the same relation to z as octant *5* does to x) and zPx (because z is in octant *2*). Thus acyclicity entails a vetoer. Allan Gibbard of the University of Michigan proved that *P*-transitive constitutions satisfying Arrow's first four axioms are oligarchic, meaning they empower a unique group of voters to impose their unanimous strict preferences on the society and grant those voters the right as individuals to veto strict collective preference counter to their own. All the lower seven constitutions represented in the illustration on the opposite page are oligarchic. To prove Gibbard's result with three voters, then, it is necessary to show only that the top three constitutions can violate *P*-transitivity. Since they differ from one another only in the naming of the voters, the argument is given only for the second of them. Because z is in preferred octant *2*, zPx. Alternative y is in octant *6*, opposite the preferred octant *4*, so that xPy. *P*-transitivity would require zPy. Yet z bears the same relation to y as octant *3* (now indifferent) does to x, contradicting the neutrality property of *P*-transitive constitutions and proving the result. The requirement of *P*-transitivity therefore entails an oligarchy. Strengthening that requirement to *R*-transitivity, which calls for weak collective preference to be transitive, eliminates oligarchies with more than a single member, leaving only the three dictatorships. The boxlike shapes in this diagram serve to locate y and z in the appropriate octants in the illustration at the left.

interests—the rule of consensus declares the two alternatives to be indifferent. A constitution that yields collective indifference whenever individuals' rankings conflict is virtually useless.

Can more attractive *P*-transitive constitutions be discovered? The philosopher Allan Gibbard of the University of Michigan has shown that they cannot. Gibbard proved that under every *P*-transitive procedure obeying Arrow's remaining axioms there exists a privileged set of individuals he called an oligarchy. This oligarchy as a group has the power to impose on the entire society its unanimous strict preference over any pair of alternatives. Moreover, each member of the oligarchy as an individual has the power to veto strict collective preference opposite to his own: whenever any oligarch strictly prefers x to y, yPx is impossible. Thus a dictator is an oligarchy of one, whereas the rule of consensus implies an oligarchy consisting of the entire society.

Not all oligarchic constitutions distribute power unequally, as the rule of consensus demonstrates. As more members are installed in the oligarchy the distribution of power becomes more nearly equal. Yet a large oligarchy increases the probability of indecisiveness, since conflicts between oligarchs imply collective indifference. Because path independence requires *P*-transitivity, Gibbard's theorem implies that the attractive property of path independence can only be purchased at the cost of indecisiveness or inequality. To escape the dilemma of choosing between the inequality of small oligarchies and the indecisiveness of large ones, collective rationality must be weakened beyond *P*-transitivity.

P-transitive constitutions satisfying Arrow's other axioms have an additional drawback. They are inflexible in the sense that, except when all members of the oligarchy are indifferent, they cannot impose different requirements for strict collective preference on different pairs of alternatives. Instead they must treat pairs of alternatives in a neutral or symmetric fashion. Hence if Tom but neither Dick nor Harry has a stake in a change from policy y to policy x, it may be appropriate to give Tom the right to veto the ranking xPy. Yet it may be undesirable to give him such power over other pairs. *P*-transitive constitutions satisfying Arrow's other axioms, however, cannot exhibit this flexibility.

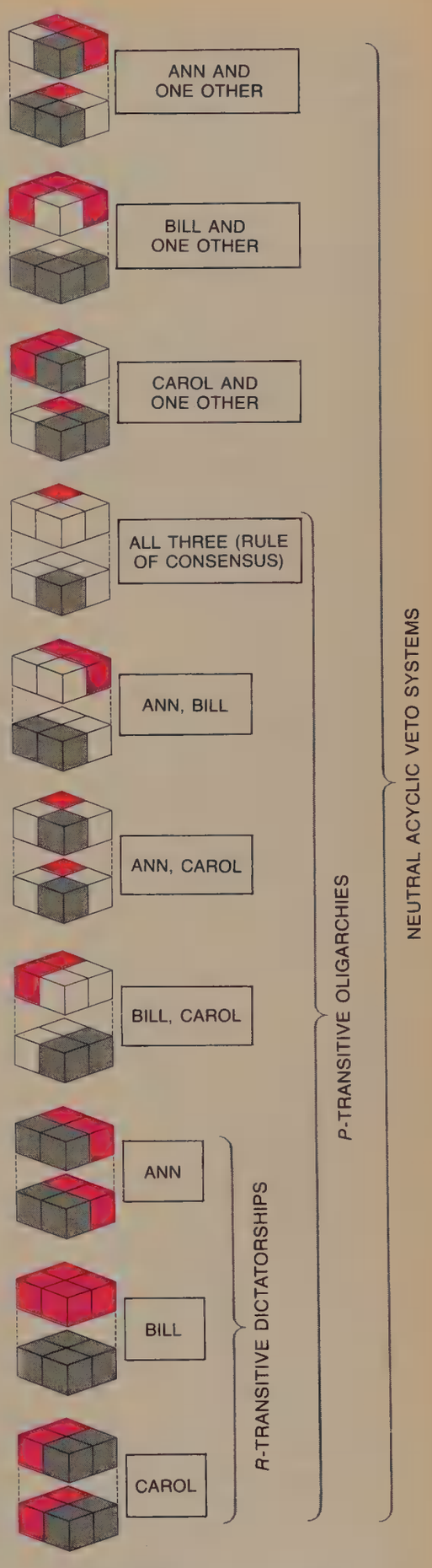
Under the oligarchic *P*-transitive constitutions satisfying Arrow's other axioms a group (or an individual) with the power to impose its will over one pair of alternatives must have the same power over all pairs. Similarly, an individual or a group with veto power over one pair of alternatives must also have that power over every pair. If a particular configuration of preferences between x and y implies that x is collectively preferred to y (Pairwise Determination guarantees that no additional information about individuals' preferences figures in determining the collective ranking of x and y), the same configuration of individuals' preferences between z and w implies that z is collectively preferred to w. It is possible to treat pairs of alternatives in asymmetric or non-neutral ways under constitutions satisfying weaker rationality conditions than *P*-transitivity.

A less restrictive requirement than *P*-transitivity is acyclicity, the absence of cycles of strict collective preference regardless of their length. For example, with three alternatives the collective preference ranking xPy, yPz, xIz is acyclic (since it has no cycles of strict preference), but it violates *P*-transitivity (since the first two rankings would require xPz). Among the collective preferences ruled out by acyclicity is the three-alternative cycle encountered in the paradox of voting.

Acyclicity is an attractive property to demand of a constitution, particularly when the procedure adopted for converting pairwise collective rankings into choices is to select the alternatives that are undefeated by all others. Acyclicity guarantees that at least one such alternative always exists in every finite feasible set, and it is the least restrictive condition that does so. In the acyclic example above, for instance, alternative x is not defeated by either of its competitors, whereas with the cyclic collective preference xPy, yPz and zPx each alternative loses to another one. Without acyclicity the appropriate collective choice is far from clear.

Acyclic constitutions satisfying Arrow's other axioms need not be neutral; they allow pairs of alternatives to be treated in an asymmetric fashion. Nonneutral procedures are quite common. In the U.S. Senate, to take one (sometimes cyclic) example, an ordinary bill passes with a simple majority, but a motion to limit debate requires three-fifths and a proposed constitutional amendment requires two-thirds. Whether or not neutrality is a desirable feature of a constitution depends on the nature of the alternatives. When candidates for office are being ranked, a neutral rule is more appealing than one favoring a particular candidate. When alternatives have asymmetric consequences, however, non-neutral rules biased against more drastic outcomes may be advantageous. The criminal-justice system provides examples at both ends of the spectrum: jury trials are not guaranteed in minor cases, and death sentences are automatically reviewed by appellate courts.

Relaxing the rationality requirement from *P*-transitivity to acyclicity makes possible many constitutions without ol-



THREE-VOTER neutral constitutions that satisfy various rationality conditions are portrayed. Each eight-octant figure illustrates the ranking of alternatives in relation to a fixed alternative x under a particular constitution. Alternatives in colored octants are preferred to x, those in gray octants are inferior to x and those in white octants are indifferent to x. The numbers in the boxes indicate which voter or group of voters is decisive in each situation.

igarchies. All the new constitutions, however, share one of the principal drawbacks of oligarchic rules. All acyclic constitutions satisfying Arrow's other axioms grant some group or individual extensive veto power.

Typical of these additional acyclic constitutions was the voting rule in the United Nations Security Council until 1965. The Security Council then consisted of five permanent members and six nonpermanent ones. A motion succeeded if it received at least seven affirmative votes and no negative vote from any permanent member. Thus each permanent member of the Security Council had a veto; no motion could pass if any permanent member of the council opposed it. These five nations did not constitute an oligarchy, however, since additional support from nonpermanent members was required for strict collective preference.

Under acyclic constitutions an individual or a group may have veto power over some pairs of alternatives but not over others. The extent of veto power under such rules must therefore be described in terms of the number of pairs over which some individual or group exercises a veto. As we have recently shown, when the number of alternatives is large with respect to the number of individuals, at least one individual must be able to veto a large number of pairs of alternatives. More precisely, as the ratio of alternatives to voters increases without limit, the proportion of pairs of alternatives over which some particular individual must have a veto approaches unity.

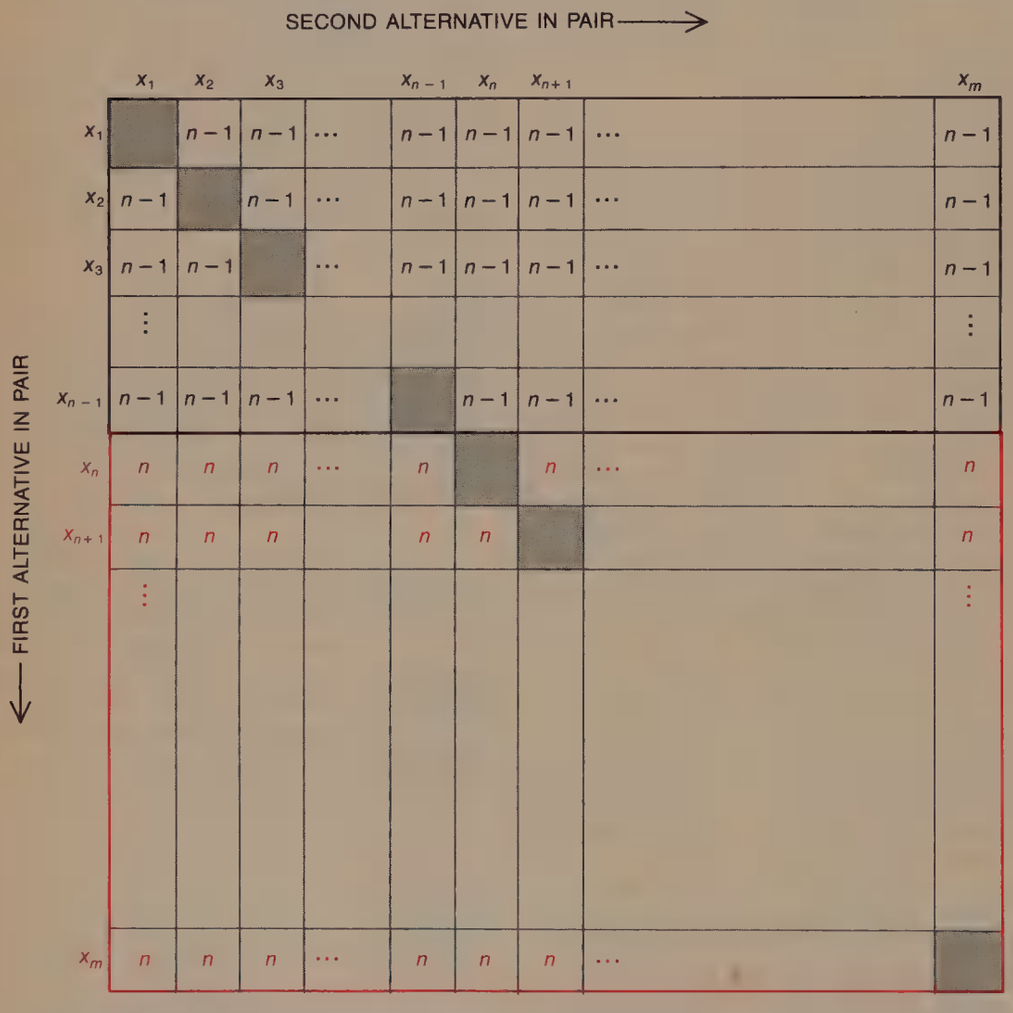
Even when there are fewer alternatives than there are individuals, their ratio is critical. *R*-transitivity and *P*-transitivity are rationality conditions applying to triples of alternatives. Acyclicity, in contrast, rules out cycles of every possible length: the absence of cycles with three alternatives does not imply the absence of cycles with four. As the number of alternatives increases, longer cycles become possible.

With only two alternatives, x and y, majority voting cannot yield a cycle. A cycle would require both that more than half of the voters prefer x to y and that more than half of the voters prefer y to x; consequently at least one individual would prefer x to y and y to x, which is clearly impossible. When there are three or more alternatives, as the paradox of voting shows, no individual need agree with all the links in the collective-preference cycle, and so the cycle does not contradict the transitivity of any individual's preference ranking.

A five-member committee further illustrates the critical role of the ratio of alternatives to individuals. Consider the constitution requiring four affirmative votes for strict preference—a constitution intermediate between majority rule and the rule of consensus. Could the cycle xPy, yPz, zPx arise under the four-fifths-majority constitution? It could not, since each individual has transitive preferences and at least four voters must agree with each link in the collective ranking. With a five-member committee at least one member would have to agree with each of the three links in the collective-preference cycle, which is impossible.

With a five-member committee and three alternatives four-fifths-majority rule has advantages over both majority rule and the rule of consensus. Unlike majority rule, no cycle can occur. Unlike the rule of consensus, no member has a veto. These advantages, however, are won at a price. The four-fifths rule is less rational than the rule of consensus: although collective preferences are acyclic, they are not *P*-transitive and so are not path-independent. The rule is less decisive than bare majority rule: although no individual has a veto, any group of two voters can block a collective ranking of two alternatives that stands opposite to their own.

Under an acyclic constitution the size of the smallest group with veto power depends on the relative numbers of alternatives and individuals. When there are only two alternatives, as the case of majority rule makes clear, the smallest veto group needed is half of the electorate. When there are as many alternatives as there are individuals, under any acyclic rule there must exist at least one individual who, standing alone, has veto



EXTENSIVE VETO POWER is characteristic of all acyclic constitutions that satisfy Arrow's first four axioms. The constitution represented here, which is defined for n voters and m alternatives, can be described in terms of the minimum number of voters whose agreement is necessary for strict collective preference between elements of an ordered pair of distinct alternatives. This number appears in the cell corresponding to the ordered pair. Unanimous consent is necessary for strict collective preference for any of the $(m-n+1)(m-1)$ pairs in the colored region of the table; hence each voter has a veto over each of these pairs under this rule. To see that this constitution cannot yield cyclic collective preferences suppose, on the contrary, it can. At most one voter disagrees with each of the links in the collective-preference cycle that lies in the black region of the table. Thus even if a different voter disagrees with each black link, some voter must agree with all of them, since at most $n-1$ pairs in the cycle lie in the region. The same individual must also agree with each colored link in the cycle. Since he agrees with every link in the cycle, his preferences are cyclic. The contradiction proves that the constitution cannot in fact cycle. When m exceeds n, this rule gives each voter veto power over $(m-n+1)(m-1)$ pairs of alternatives. When the ratio of m to n is large, the rule is unattractive because it is highly indecisive. Yet as the authors have recently demonstrated, every acyclic rule satisfying Arrow's first four axioms gives someone veto power over at least this number of pairs when m exceeds n. Even though the rule is unattractive, a better one cannot be found in the circumstances.

power over some pairs of alternatives. In intermediate cases, as the constitution stipulating a majority of four-fifths illustrates, some minority groups have veto power. Constitutions that grant veto power to a large number of small groups are likely to lead to deadlock rather than to decision.

An uncompromising egalitarian would argue that since some individuals must have veto power, all individuals should have it. The larger the set of vetoers, however, the more frequent the incidence of collective indifference, since indifference occurs whenever two individuals with veto power in appropriate directions rank x and y in opposite ways. The palatability of acyclic constitutions thus depends on the ratio of alternatives to individuals. When the number of alternatives is only slightly smaller than the number of individuals, "small" groups must be endowed with extensive veto power, although there need not be an individual with veto power. As the number of alternatives increases, the size of the smallest vetoing group required for acyclicity grows smaller. With as many alternatives as there are individuals, at least one individual must be able to veto some pairs. As the number of alternatives increases further, the proportion of pairs over which the individual has veto power approaches unity.

The "impossibility theorems" that began with Arrow's famous proposition define constraints on a society's choice of a rule for collective decision making. The constraints are severe. Three widely shared objectives—collective rationality, decisiveness and equality of power—stand in irreconcilable conflict. If society forgoes collective rationality, thereby accepting the necessary arbitrariness and manipulability of irrational procedures, majority rule is likely to be the choice because it attains the remaining goals. If society insists on retaining a degree of collective rationality, it can achieve equality by adopting the rule of consensus, but only at the price of extreme indecisiveness. Society can increase decisiveness by concentrating veto power in progressively fewer hands; the most decisive rule, dictatorship, is also the least egalitarian.

There is little comfort here for those designing ideal procedures for collective choice. Nevertheless, every society must make collective choices and devise voting procedures, however imperfect they may be. Axiomatic analysis, pursuing the line of investigation begun by Arrow 33 years ago, has yielded a richer understanding of existing voting methods and may eventually yield better ones. It also demonstrates that the opportunities for improvement are severely limited. Stark compromises are inevitable.

BIG BOOM FOR REAL ESTATE INVESTORS.

Thousands of people like you find new opportunities for income, appreciation and tax advantages through the CENTURY 21® network.

Tax laws have changed. And now real estate is an even better investment than ever before.

At CENTURY 21 offices all over the country, we're showing people how to take advantage of the latest tax laws. How they can use leverage. How they can now shelter more of their income. Even income from salary, stock dividends and other sources.

A CENTURY 21 Investment Professional can show you, too. Through real estate, you may be able to get a greater annual cash flow, after taxes, than through other traditional types of investment. And you don't have to invest a lot of time and effort. Because a CENTURY 21 office can also handle the details of property management for you.

The following chart illustrates the potential advantages of real estate over other investments:

	REAL ESTATE	MONEY MARKETS	STOCKS
Cash Flow	✓	✓	✓
Depreciation	✓		
Loan Amortization	✓		
Appreciation	✓		✓

The opportunities are impressive. And that's one of the reasons investment business is booming at CENTURY 21 offices. Last year, we helped investors buy or sell real estate worth over $4 billion. We're North America's leading real estate sales organization. And now, we're leading the way in real estate investment, too.

Century 21®

Call this toll-free number for a CENTURY 21 Investment Professional.
1-800-228-3399

Nebraska Residents please call 1-800-642-8788. In Alaska and Hawaii call 1-800-862-1100.

©1983 Century 21 Real Estate Corporation, as trustee for the NAF ® and TM-trademarks of Century 21 Real Estate Corporation Printed in U.S.A. Equal Housing Opportunity

EACH OFFICE IS INDEPENDENTLY OWNED AND OPERATED.

The Stave Churches of Norway

In the 10th century Norwegian builders blended pagan and Christian elements in wood churches. Some are still standing, demonstrating that with sound design and maintenance wood buildings can be permanent

by Petter Aune, Ronald L. Sack and Arne Selberg

There is ample evidence that in antiquity large structures were built not only of stone, brick and mortar but also of wood. Yet virtually all the large structures that have come down to us from antiquity and even from medieval times were built of stone. The reason is almost self-evident: stone, brick and mortar are far from indestructible but under most conditions they are more durable than wood, which is vulnerable to decay, pests, fire and neglect. An exception is the stave church of Norway, a type of wood structure that dates back to the 10th century. The exception is a meaningful one.

The staves of the stave churches are stout wood columns. "Stave church" is translated from the Norwegian *stavkirke,* and *stav* is related both to the stave of a barrel and to a wood staff. Stave churches were built in large numbers in Norway from the 11th century to the 14th. In the construction of the churches pagan and Christian elements were combined. The result is a design of primitive grandeur, with an exotic exterior and a serene interior that is distinctly human in proportion. The most intriguing aspect of the stave church, however, is the durability of its structure. How were these wood buildings made so that some of them could have survived for more than 800 years?

We are three structural engineers who have surveyed more than half of the stave churches still in existence. We have found that there are three reasons the buildings have been able to last so long. First, extraordinary care was taken in the preparation of the wood; the unique seasoning methods of the early Norwegian builders yielded materials that are superior to modern lumber. Second, the overall design is an excellent one, well suited to the exposed sites on which many of the churches were built. The structure is capable of sustaining the vertical forces generated not only by the mass of the church but also by heavy loads of snow. In addition it provides the combination of strength and stiffness that is needed to withstand the lateral forces exerted by high winds. Third, small but significant structural innovations have served to protect the wood from deterioration. The lesson is that with appropriate design and with special care taken in preparing the building material large wood structures can be permanent. Indeed, the history of the stave church suggests that the frequent assumption that wood is suitable mainly for small buildings or for buildings not expected to have a long life might well be reexamined.

The stave church incorporates two main types of structural assemblies made out of wood: truss structures and beam structures. In an ideal truss structure each member is subject either to tensile stress or to compressive stress but not simultaneously to both. Furthermore, there are no other types of stress on the members. A simple example of a truss structure is a triangular assembly with the apex of the triangle at the top; such a structure might be found in a roof. A downward vertical force applied at the peak puts the two diagonal members into compression and the bottom member into tension. Such a downward force could be supplied by the weight of the roof or by a load of snow.

In a beam structure, on the other hand, tensile and compressive stresses can both be present in a particular cross section of one member. There can also be other kinds of stress in the structure, such as those exerted by shear or bending. A simple example of a beam structure is an assembly with a horizontal beam attached to the upper end of two vertical columns. Such an assembly could form part of the frame for a house. The weight of the top beam and the loads it supports provide a downward force that compresses the vertical members. The vertical members, however, tend to curve outward; as a result they are subject not only to compressive forces but also to bending stresses.

Beam structures go back to prehistory. The design can be executed simply by laying a horizontal beam on two supports. The truss design, however, also has a long history. Roman architects are remembered primarily for their work in brick, mortar and marble, but they also did superb work in wood. Trajan's Column in the Imperial Forum of Rome was erected in A.D. 114 to commemorate the emperor's conquest of the Dacians, who lived in what is now Romania. Among the figures on the column is a group showing that in A.D. 104 Apollodorus of Damascus, the imperial architect, built a wood bridge across the Danube; the design of the bridge incorporated trusses. The way truss elements and beam elements are combined in the stave church does much to explain why the churches are so sturdy.

Wood continued to be employed as a building material in Europe after the collapse of the Roman empire. The fabled ships built by the Vikings demonstrate that techniques of construction in wood were highly developed in the Nordic countries by the early Middle Ages. Moreover, the Viking ships kept the Nordic peoples in contact with the centers of culture and hence made available to them the building-construction technology that had evolved elsewhere in Europe.

FANTOFT STAVE CHURCH is a well-preserved example of the early stave churches. Built in about 1200 at Fortun, a village at the head of Sogne Fjord, it was dismantled and reconstructed in 1883 at Fantoft near Bergen. The stave church is named for the wood columns that are its primary supports. In the exterior of the church the builders combined pagan and Christian motifs. The dragon heads on the gables, similar to those on the prow of a Viking ship, overshadow the crosses below. The building consists of a rectangular nave (*in the foreground*) and a smaller rectangular chancel (*at the rear*). The Fantoft church has a basilican design: the central nave is raised above the aisles on each side. Of the three large roofs, the highest covers the nave. The middle roof covers the aisles. The lowest roof covers a walkway called the ambulatory.



The influence of such technology in Norway was strengthened by the introduction of Christianity. The Christian religion was slower to reach Norway than it was to reach most of the rest of Europe. The first king to attempt to bring Christianity to Norway was Haakon the Good, who ruled from A.D. 935 to 961. Haakon had three churches built in the Møre district. The churches were soon burned and the priests whom Haakon had imported from England were killed. The missionary efforts of several later kings were also fiercely resisted. It was not until the reign of Olav Kyrre, who ruled from 1066 to 1093, that a lasting official Norwegian Church was established. Olav set up episcopal sees at Trondheim, Bergen and Oslo, and he probably also instituted a large program of church building.

Thus in the 10th and 11th centuries the number of churches in Norway grew, albeit not without setbacks. The early missionary churches were simple buildings consisting of a large rectangular nave with a small chancel attached to its east end. The nave provided room for the congregation, whose members generally stood during the service; the chancel was for the altar and clergy.



GOL STAVE CHURCH was built between 1200 and 1250. In 1884 it was moved to the Norwegian Folk Museum at Bygdø near Oslo. The view in the photograph is from the nave into the chancel, where the altar stands. The Gol church also has a basilican design. The central nave is defined by the interior staves: 14 wood columns eight meters high. Six of the staves have been terminated above the floor. The curved brackets under the shortened staves allow the loads to be transferred to the full-length corner staves. The corner staves transmit the loads to the substructure. The stave assembly is given rigidity by sets of curved braces called arcading (visible above the head of the crucifix) and by the intersecting diagonal struts called St. Andrew's crosses (visible behind the body of the crucifix). The staves visible in the rear support the chancel.

Some of the missionary churches of Norway have been excavated in modern times. The remains of structural members found buried at the sites suggest that many of the nave-and-chancel churches were built with the ancient method of palisade construction. This method called for hewing tree trunks in half lengthwise and sinking the timbers into the ground next to each other to form the walls. In the refinement known as posthole construction the corner columns were sunk into the ground and horizontal foundation beams were laid between them resting on the ground. Wall planking was erected on the foundation beams.

Posthole construction was an advance over the palisade method, but the foundation members, being in contact with the ground, remained susceptible to rotting. That is why few specimens of posthole buildings survive, even though archaeological excavations show they were common throughout Europe. A further innovation extended the life of the foundation beams and made larger structures possible, namely the building of a complete foundation above ground level. On a layer of flat stones was set a rectangular frame of heavy members called sill beams; the main vertical columns were put at the corners of the frame where the sill beams intersected.

Vertical wall planks were inserted into a groove in the sill beam. In some instances the narrow space between planks was closed with a tongue-and-groove joint; in others it was covered with a thin wood spline. Although it is the larger corner posts that are called staves, the vertical wall planking does call to mind the staves of a barrel. The stave style, with its upright wall planks, stands in contrast to the commoner method of wood construction in Norway, which is called *laft* construction. In *laft* construction the wall members are laid horizontally and are held together at the corners by interlocking notches.

After the first stave churches were built in the middle of the 10th century, the design was intensively refined over a period of several centuries. The church at Urnes on Luster Fjord, a branch of Sogne Fjord, provides a convenient illustration of the process of development. Built in about 1150, it is thought to be the oldest existing stave church in Norway. Findings made in an excavation under the church floor in 1956 suggest, however, that two earlier churches had been built on the same site.

The first Urnes church was probably put up in the time of the transition to Christianity. It was built in the palisade style: all the structural members were embedded in the ground. The sec-

ond church was put up in the latter half of the 11th century. It was built by the posthole method. Both of the early churches were small structures of a nave-and-chancel design.

In the middle of the 12th century the second Urnes church was torn down. Some of its parts were utilized for the building of the third church, the one that is still standing. Thus three churches were built at Urnes in less than a century and a half. The design of the third church was a dramatic departure from that of its predecessors. The central section of the nave is raised above the rest of the building on a framework consisting of 16 large staves, which define the nave and the surrounding aisles. The roof of the central compartment is about two meters higher than that of the aisles. The striking design of the third Urnes church spread throughout Norway and was followed in the construction of many later stave churches.

A significant influence on the new design appears to have been the cathedrals in the basilica style that were going up elsewhere in Europe at the time. Most of the basilican cathedrals were built of stone, but it is possible the Norwegian builders saw ways of adapting the design to wood. Most basilican cathedrals have at one end a large semicircular apse, a structure adjacent to the chancel. The large rectangular nave is separated from the two side aisles by two parallel rows of columns. The builders of the stave churches had examples of the basilican design close at hand: stone cathedrals of this type were built at Oslo, Hamar, Stavanger, Bergen and Trondheim. Because of its similarity to the basilican cathedral, we shall refer to the kind of stave church that first appeared at Urnes as the basilican stave church.

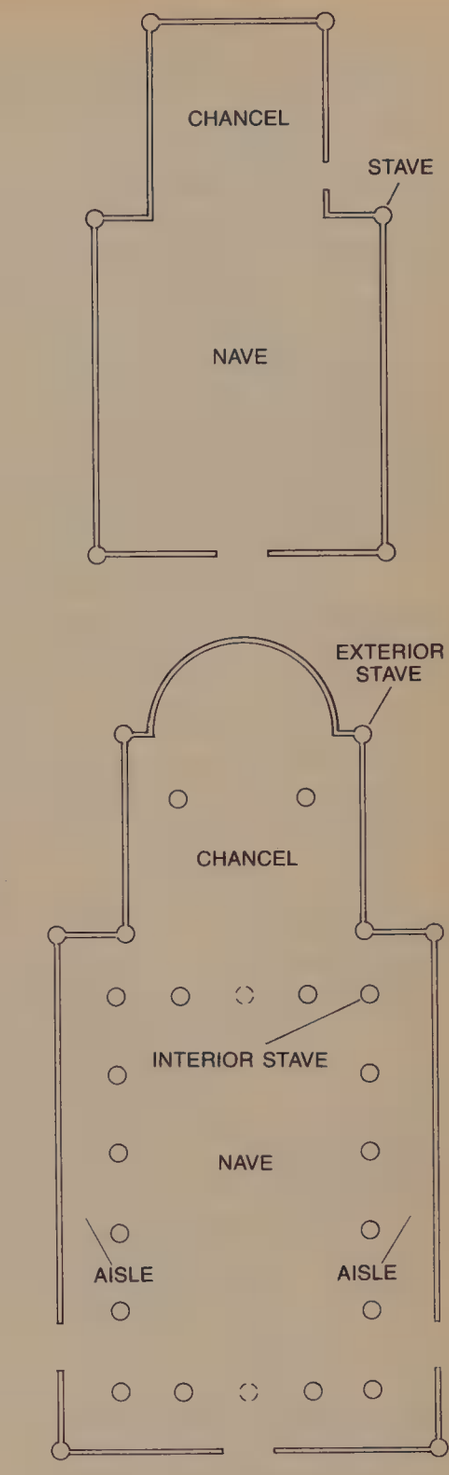
Although the basilican cathedrals could well have been the decisive influence on the builders of the church at Urnes, other models were culturally even closer at hand. Kenneth J. Conant of Harvard University notes that excavations at Gamla Uppsala in Sweden have revealed the remains of a pagan temple with a plan much like that of the basilican stave church. According to Conant, in the temple "there was a square central compartment with corner timbers more than two-thirds of a meter in diameter and a smaller post between [each pair of larger timbers]. The outer wall was supported by light posts, and [was] relatively low, while the large corner timbers indicate a towerlike proportion for the central square." The Gamla Uppsala temple was built in the middle of the 11th century.

Moreover, excavations at Lund in Sweden have uncovered the remains of churches dating from about the same time. These early churches were constructed on the palisade design. Some of them also had freestanding interior staves, as the basilican stave churches do. Such old missionary churches are generally regarded as the forerunners of the stave churches.

Hence in converting from the simple nave-and-chancel church to the grander basilican design the builders in the Norwegian villages probably drew on both Christian and pagan sources. The conclusion is reinforced by the decorations with pre-Christian motifs that appear on the outside of stave churches. In many of the structures the exterior is adorned with elaborate carvings of dragons and other zoomorphic figures in an exuberant tangle. The designs have been traced to several possible sources, including the carved runic stones of Scandinavia and certain kinds of manuscript illumination then being done in Ireland. Whatever the source of the carvings, they have a distinctly pagan spirit. Only after 1100 did Christian symbols such as plant tendrils begin to appear among the dragons. In addition to the zoomorphic carvings the stave churches have on their gabled roof dragon heads much like those on the prow of a Viking ship. They are often larger and more dramatic than the crosses on the roof.

The stave design was clearly popular with the Norwegian villagers. It has been estimated that in A.D. 1300 there were about 1,000 stave churches in the country. Many of them were built in the area north of Bergen along the tributaries of the Sogne Fjord, but there were also stave churches from the south coast up to Nidaros (the modern Trondheim). Moreover, although there are no surviving stave churches in Sweden and Denmark, recent excavations suggest that such churches were also built there.

As is customary with Christian ecclesiastical buildings, most stave churches are built with the long axis of the nave laid out from east to west and the chan-

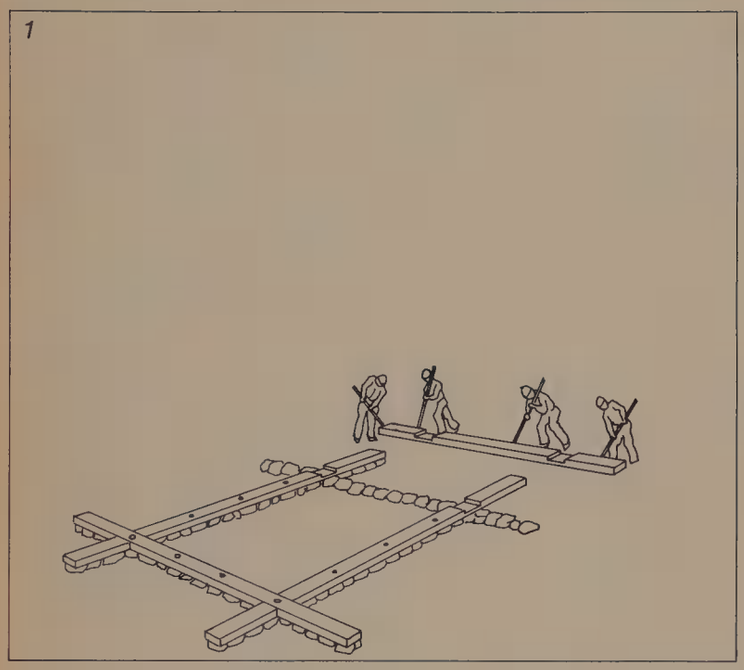


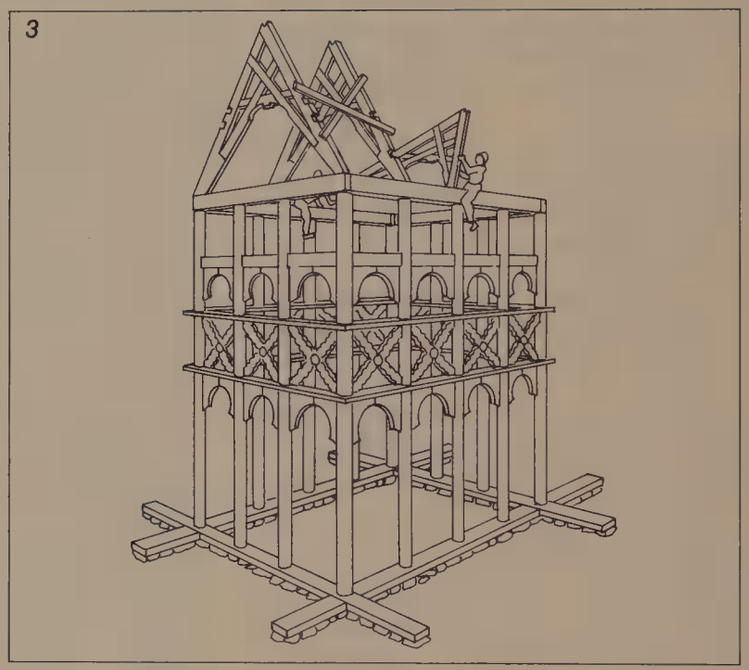
FLOOR PLANS show the development of the basilican stave church, the first of which was built in about 1100. Before the 12th century stave churches had a simple nave-and-chancel design (*top*). Large horizontal sill beams laid on a foundation of flat stones made up the footings. At the corners were six large staves. The vertical wall planks between the corner staves were inserted in the sill beam at their base. In the basilican churches the interior staves defining the raised central compartment are laid on four large raft beams that form a rectangle (*middle*). One influence on the basilican church was the design of pagan temples such as the one at Gamla Uppsala in Sweden (*bottom*). The plan of the temple has been reconstructed from the few column supports that were excavated intact (*circles*). The temple had low outer walls and a towerlike central compartment. The central compartment was supported at the corners by columns more than two-thirds of a meter in diameter.

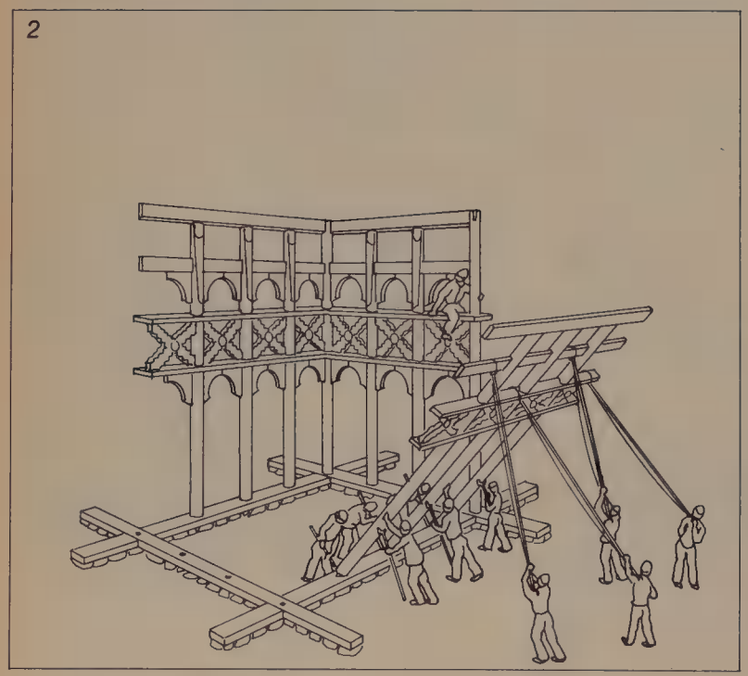
cel at the east end. The churches were generally built on visually prominent sites and are therefore exposed to the full effect of sun, wind, rain and snow. In order to construct a tall wood building on such a site several significant engineering problems must be overcome. The foundation must provide a stable footing for the building and protect the lower members from rot. The vertical load imposed by snow and the mass of the structure must be conveyed to the footings in such a way that the building is stable. The structure must be able to sustain the substantial lateral forces exerted by high winds. Provision must be made for protecting exposed joints from deterioration and for replacing parts that are subject to rot. How these problems were solved emerges from a survey we have done of 16 of the surviving stave churches. The basilican stave churches are the most complex of the group, and in what follows we shall be concentrating on them.

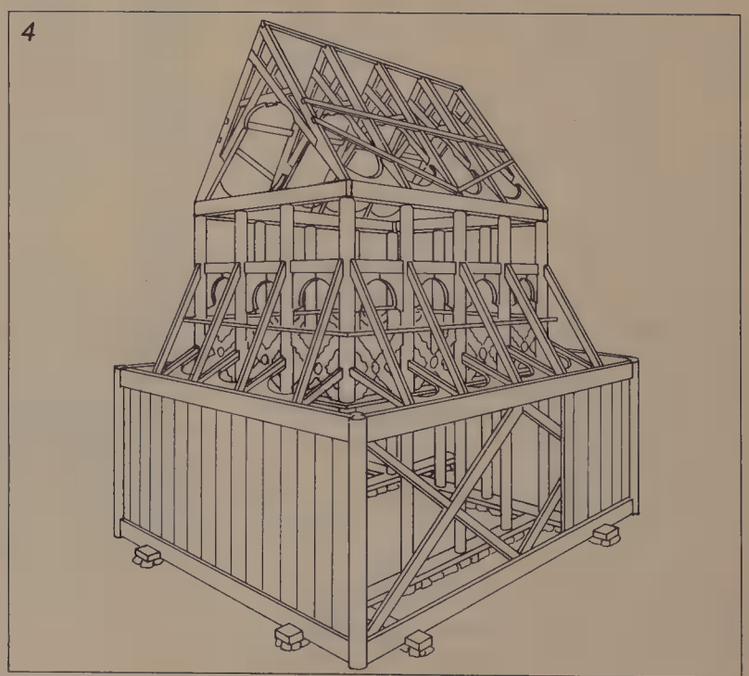
The main interior staves in the basilican stave church are supported by four horizontal members called raft beams. The raft beams are laid on rows of flat stones set on the ground and are joined at the corners by mortises (meshing notches). They are arranged in two parallel pairs that intersect to form a square with the beams extending beyond the corners in an arrangement much like the symbol designating a number (#). The beams are notched at their outer ends to accommodate the sill beams. Depending on the exact design of the church, other foundation beams can be added to the interlocking grid of raft beams.

The foundations of the nave and the chancel are often constructed so that they function independently of each other. The interaction of structural components resulting from this system makes it possible for forces to be transmitted through the building in a complex way. The configuration has the distinct disadvantage of introducing large stresses into the structure if one of the units settles into the soil more than the other. Our survey shows, however, that there is almost no such "settlement distress" in the stave churches. Therefore it









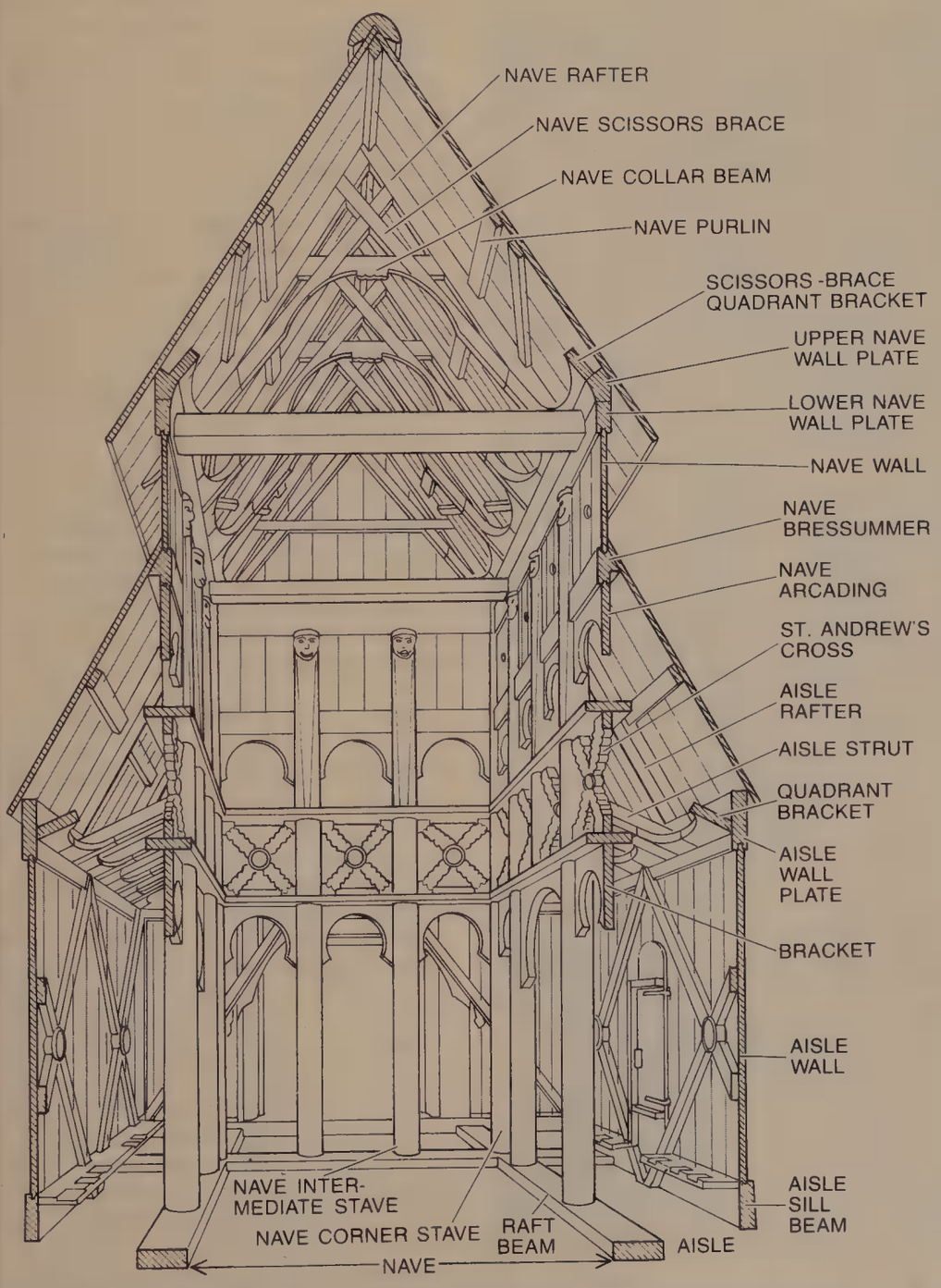
CONSTRUCTION OF A SMALL STAVE CHURCH with the basilican design was carried out with methods much like those employed to erect a small frame house. The raft beams were laid on a bed of flat stones and joined by mortises (*1*). Each assembly of staves corresponding to one face of the central compartment was raised on the raft beam and held in place by temporary bracing (*2*). The tenon on the bottom of the stave was tapered so that it would slide easily into the mortise in the raft beam. The assemblies were joined at the corners by the arcading. When the four sides of the central compartment were up, the roof was put on (*3*). The aisles were then added (*4*). The

appears that the soil conditions in the areas where the stave churches survive are favorable to the design.

The primary structural members that transfer the load from the roof and walls to the footings are the staves. The interior staves can be as long as 11 meters; their diameter varies from 30 to 40 centimeters. A tenon, the projection that fits into a mortise, connects the base of the stave to a corresponding mortise in the raft beam. At the top of the stave is a structure called a wall plate, made up of two smaller beams; the rectangle composed of the nave wall plates provides the base for the roof of the church. Thus the central compartment of a basilican stave church has the form of a parallelepiped. The top of the parallelepiped is defined by the nave wall plates, the bottom by the raft beams. The four sides of the compartment are defined by the interior staves, and there are generally from three to five staves to a side.

The rectangular central compartment must be braced to keep it from collapsing. Rigidity in several directions is required. Each side must be braced to prevent the staves from falling over like a row of dominoes. Resistance to forces in the plane of the staves is provided partly by horizontal beams called bressummers, which are placed from 1.25 to two meters below the top of the stave. As we shall see, the bressummer has several significant functions in addition to bracing the staves. In the right angle between the bressummer and the stave is a curved wood bracket that supplies additional stiffness; the brackets are collectively termed arcading.

Near the arcading in many stave churches is a bracing assembly made up of two intersecting diagonal members between each pair of staves. Such assemblies, called St. Andrew's crosses, first appeared in about 1200. Many of the churches put up before then were fitted with crosses long after they had been built. The crosses were probably intended to give horizontal strength to the assembled staves; in many churches, however, the crosses are too loose to provide much resistance. Indeed, climbing in the interior of the Torpo stave church, one of us (Sack) narrowly escaped injury when a cross almost pulled loose from its moorings. On some very long staves a second tier of arcading has been put below the crosses. The brackets and crosses are fastened to the staves and to the horizontal beams by round wood pegs about 25 millimeters in diameter. Some of the pegs have an enlarged head, but in many instances the peg was fixed in place by a wood wedge driven into its outer end.



planks making up the aisle wall were inserted into the sill beam at their base. The completed church is shown at the right. The drawing is based on the church at Borgund, which was built in about 1200 and is one of the most beautiful and best-preserved of the stave churches. The view is from the chancel through the nave toward the main entrance. The central compartment is defined by 14 interior staves 6.7 meters high.

As the builders of the stave churches gained experience they began to change the configuration of the interior staves in order to open up the nave, allow easier entry and provide a better line of sight from nave to chancel. This was done by shortening some of the staves. The staves at the corner of the central compartment were left intact but some staves between the corners were cut off about halfway up, so that the bottom of the staves rested not on the footings but on a wood arch formed by a pair of brackets. This feature was introduced with increasing boldness. In some early churches one stave at the west end and one at the east end were shortened. In the church at Heddal, which was built in about 1300, short interior staves alternate with full-length ones. In the church at Hurum and the one in Lomen only the corner staves continue to ground level.

Such modifications were probably possible only after much experience with stave construction. Shortening the interior staves calls for a sure sense of the strength of the structure. Moreover, the modification makes it necessary to have a large set of brackets to carry the load from the shortened staves to the corners where the intact staves convey the load to the footings.

The builders probably raised the inner compartment of a small stave church the way a small frame house is erected today. Each assembly of three to five staves corresponding to one wall of the central unit was put together on the ground at the building site. After the raft beams had been laid down the tenon at the bottom of the stave was inserted into

the mortise in the raft beam; the tenon was tapered so that it would slide easily into the mortise when the stave was lifted. The flat assembly was then raised and held up by temporary supports.

The process was repeated until all four sides of the central compartment were up. The stave assemblies were connected by the curved corner brackets. In addition each corner was generally fitted at the level of the bressummer with a horizontal "knee bracket," which gave the central unit strength to resist horizontal deformations. The central compartment of the stave church is geometrically similar to a cardboard box with the top and bottom removed; very strong winds blowing at a skew angle to the church could flatten the compartment much as cardboard boxes are flattened before being stacked in a supermarket. The knee brackets contribute resistance to such flattening.

When the central compartment was up, the roof was put on. The roof of the stave church is sharply peaked, generally with a rise of 3 : 2, or an angle of 56 degrees. The rafters are braced by a pair of beams; the beams cross under the peak of the rafters. Such a "scissors brace" often has a rise of 1 : 1, or an angle of 45 degrees. A horizontal beam called a collar beam is included in the roof bracing near midheight; it is generally connected to the rafters and to the scissors braces. Light secondary members called purlins, which run parallel to the roof ridge, transfer roof loads to the rafters and the roof bracing. In the later stave churches diagonal bracing was incorporated into the plane of the roof to keep the structural components from collapsing.

The roof assembly of rafters, scissors brace and collar beam resembles an orthodox truss structure but does not function as an ideal truss. An assembly found in many European wood buildings is the German truss, or scissors truss. It consists of nine subunits: four diagonal members at the top, two lower ties, two interior struts and a vertical member [*see bottom illustration on opposite page*]. As in other truss structures, each member is either in tension or in compression: under a downward vertical load the four top members and the two interior struts are in compression whereas the two lower ties and the vertical member are in tension.

In the scissors braces of the stave church roof, however, the vertical member is conspicuously absent. In its place is the collar beam. As a result the brace does not operate as an ideal truss; an analysis of the structure shows that many of the beams are subject to bending forces in addition to tension and compression. It is not surprising that the 12th-century Norwegian builders did not construct an ideal truss. Truss design was not formalized until the beginning of the 16th century, and before that time local solutions to engineering problems were worked out by trial and error. The collar beam nonetheless serves an important function: it redistributes the forces in the assembly and thereby minimizes the horizontal components of the forces delivered to the walls. If the collar beam were absent, the tops of the walls could be pushed apart under a heavy snow load.



PAGAN IMAGERY dominates the carvings found on the exterior of many stave churches, as can be seen in this section of the carving from the church at Hopperstad, which was built in about 1180. The zoomorphic designs in the panel are related to carvings on Scandinavian runic stones and to Irish manuscript illumination. The motifs include the dragon and the griffin. Christian motifs such as plant tendrils were inserted among the pre-Christian designs.

Once the rafters and the roof bracing were up light wood planks were added running down from the ridge to the eaves. Exterior planks were then laid parallel to the roof ridge; the two layers of planking were usually fastened together with wood pins. In some of the churches the two layers were covered by a third layer made up of wood shakes, or rough shingles. The shakes were often cut in an unusual form that gives the roof the appearance of fish scales.

When the roof was up, the aisles were added. The roof of the aisles, being lower than that of the raised central compartment, emphasizes the vertical lines of the building and gives the exterior profile the look of a set of inclined steps. The upper end of the sloping rafters that support the aisle roof is attached to the interior stave with wood pegs. The lower end rests in a notch cut in a horizontal beam called the aisle wall plate, which is at the top of the aisle wall.

Under the aisle rafters is a crossbeam with an incline less steep than that of

the rafter. Called the aisle strut, it is attached with pegs to both the aisle wall and the interior stave. It can therefore transmit both compressive and tensile forces. In contrast, the aisle rafter is attached only to the interior stave. At the aisle rafter's other end, where it meets the wall plate, it merely rests in a notch. It can therefore transmit only compressive forces.

Between the aisle struts is a large curved brace known as a quadrant bracket. Connected to the aisle wall plate, it stiffens the entire aisle structure considerably. It is particularly effective in resisting horizontal forces parallel to the long axis of the nave.

One of the final steps in construction was to fit the vertical wall planks between the sill beam and the aisle wall plate. The sill beam and the wall plate were grooved to accommodate the tenons that project from the top and bottom of the wall planks. The gap between a pair of planks was sealed with a tongue-and-groove joint or a spline.

The aisle assemblies that flank the nave make a notable contribution to the lateral strength of the entire church. Consider a wind blowing at a right angle to the long axis of the nave. The forces on the windward aisle wall are transmitted from the planking to the aisle wall plates and the interior stave through the aisle strut and the quadrant brackets. The structure will tend to deform in the direction of the wind. In the deformation the windward aisle rafter tends to respond as a tensile member. Since the rafter is attached at one end only, it is ineffective in transferring the load.

The lateral forces are transmitted across the central compartment of the church by the interior staves, the roof structure, the arcading and the St. Andrew's crosses. On the lee side of the building the aisle rafter and the aisle strut act as a small truss to transfer the load to the aisle wall and the exterior staves. Thus the aisle assemblies fulfill a stabilizing function much like that of the flying buttress in a Gothic cathedral. In the Gothic cathedral, however, it is the great mass of the buttress that provides stability. In the stave church it is the capacity of the entire cross section to transfer forces efficiently from one side of the building to the ground on the other side that is the stabilizing factor.

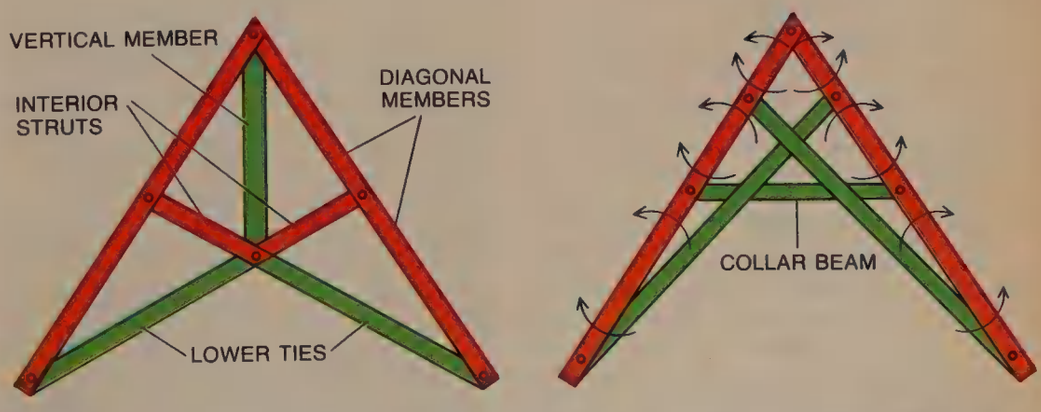
The builders of the stave church relied on an interesting combination of beam members and truss structures to provide resistance to vertical and horizontal loads. The roof is a modified truss assembly. The interior compartment, which must resist both vertical and horizontal forces, is made up primarily of beam members. The aisles, which offer crucial lateral resistance, act as small trusses. The combination of truss and beam components is highly effective.



DRAIN IN A SILL BEAM helps to protect the vulnerable footings of the stave church against rot. At the left in this view of the footings of the church at Øye is the bottom of an exterior corner stave. The large horizontal member is the sill beam; above it are the aisle-wall planks. Water running down the wall can collect in the joint between the planking and the sill beam. To drain the joint the builders put small holes every few meters along the sill beam. The outlet of one such drain appears at the right. As a result of this feature the footings in most stave churches are in excellent condition. The churches are also painted with tar every two to five years.

The materials employed in building the stave church are almost as significant as the structure itself. In the selection and preparation of wood the church builders appear to have been extraordinary craftsmen. The staves, wall planks and many other structural components are Scotch pine (*Pinus sylvestris*), a tree that is abundant in Norway. Great care was taken in selecting and seasoning the large main staves. From the many available trees the builders selected a few from which the tops were removed while the tree was still standing. The tree was left on the stump to dry for five to eight years before it was felled and the large structural members were cut to size. The outer sapwood of the tree was taken off; only the durable inner heartwood went for the staves. The wood for many of the curved brackets was from a Norwegian species of birch (*Betula verrucosa*). The wood for many of the pins and other connectors was common juniper (*Juniperus communis*), a dense softwood.

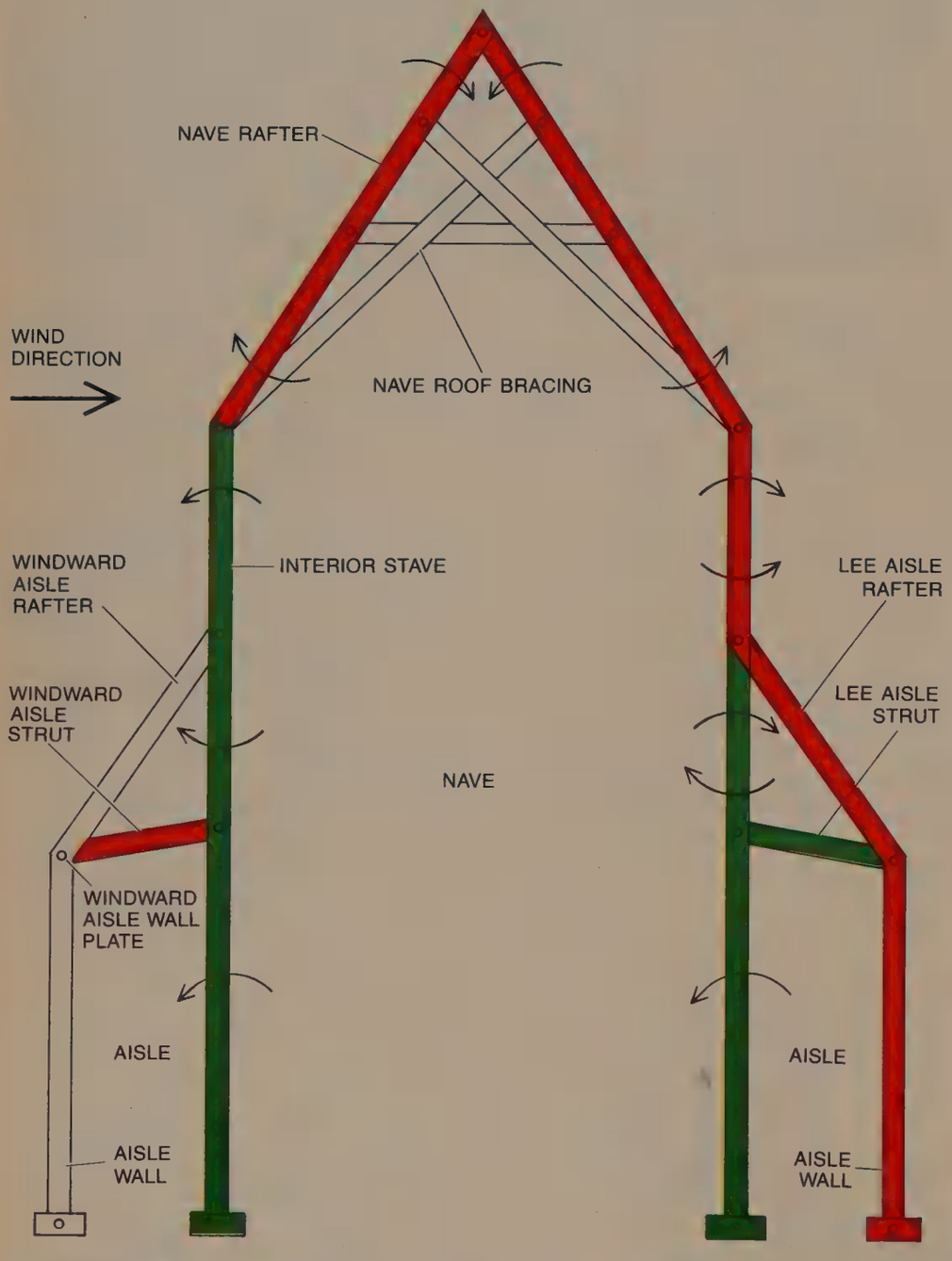
Our survey of the stave churches shows that the seasoning practices of the



ROOF BRACING of the stave church is a modified truss structure that functions differently from an ideal truss. In an ideal truss structure each member is either in tension (*green*) or in compression (*red*) but not in both at once. There are no other kinds of force in the assembly. At the left is the structure known as the German truss, which operates as an ideal truss. In the scissors bracing of the stave church roof, shown at the right, the vertical member is replaced by a horizontal collar beam. As a result some of the members are subject to bending forces (*arrows*). The collar beam helps the structure resist the lateral force exerted by a heavy snow load.

medieval builders yielded lumber that is superior to what is available to builders today. In most of the 16 churches we visited the tall interior staves were in remarkably good condition. In the church at Urnes one 800-year-old stave was still exuding pitch. The wood of the original staves was sound and relatively free of longitudinal cracks. Indeed, the worst case of longitudinal cracking was in two interior staves that had been installed about 40 years ago; both had cracks some 40 millimeters wide running from top to bottom. As is to be expected, the exterior staves had generally deteriorated more than the interior ones, but even the exposed staves appeared fundamentally sound.

The stave design helped the churches to withstand fire. Large wood members such as the primary structural staves can survive a fire because it takes time for them to burn through. Thus a stave church could last through a fire without collapsing, and when it was repaired, the charred members could be replaced. If the fire were intense enough, however, even the large structural members would burn through. A number of stave churches are known to have been destroyed by fire. A program is under way to protect all the remaining stave churches with sprinkler systems.



RESISTANCE TO LATERAL FORCES in the stave church is provided by the entire cross section of the church, which functions as a single assembly. In this drawing of a transverse section across the long axis of the nave the structural members in tension are indicated in green and the members in compression are indicated in red. Bending forces are designated by arrows. A strong wind is blowing from left to right. In such a wind the aisle wall plate on the windward side transfers the load to the interior staves by means of the quadrant bracket and the aisle strut. (In the computer simulation on which the drawing is based the arcading was omitted.) The staves, the roof structure and the roof bracing transfer the load across the central compartment. On the lee side the aisle rafter and the aisle strut function as a small truss that shifts the load to the leeward aisle wall. Thus the aisles have a stabilizing function like that of the flying buttress in a Gothic cathedral. In the Gothic cathedral stability is provided, however, by the mass of the buttress. In the stave church it is provided by the response of the entire cross section.

The resistance of the stave churches to the ravages of climate and time has been aided by a number of structural details. Along the bressummer and the outside of the sill beam the potential for rotting is great. Water can run down the vertical wall planks until it is stopped by the horizontal beam, where it can collect and penetrate the joint. To counteract the problem the builders made small drain holes at intervals of a few meters along the bressummer and the sill beams to allow water to escape from the groove in which the wall planks rest. The strategy has been effective: we noted only one sill beam in an advanced state of decay.

In many of the churches a covered walkway called the ambulatory runs around the building. The ambulatory does not serve any primary structural function, but it shields the lower part of the church from the weather. In one church bulbous wood covers were put around the base of the staves. The coverings are replaceable; if rot sets in, new ones can be installed. Most of the churches are to this day painted with tar every two to five years, a practice that gives them a characteristic smell.

Two factors of a more general nature also contribute to the durability of the stave churches. First, many of the structural members are stressed at a very low level. In most buildings put up today the supporting members are under a stress that corresponds to a large fraction of their capacity to sustain stress. This is to save money and to reduce weight and cross section. Our calculations show, however, that in the stave church the staves are under a stress that is only about a tenth of their capacity. Such a great reserve means that it is unlikely the members will fail under overload conditions. Second, the detailed shaping of all the pins and connecting elements in the building was quite accurate. As we have noted, the capacity of the cross section of the stave church to transfer loads laterally is a significant element in its overall structural integrity. Only if the grooving, notching and shaping of all the connecting members is accurate is the load transferred efficiently with no distress to the building.

Over the centuries the stave churches have been more successful in withstanding decay and structural failure than they have been in meeting another

threat: the changing needs and tastes of the communities in which they were built. In about 1350, when the number of stave churches was at its highest, the bubonic plague struck Norway, as it did much of the rest of Europe. The population of the country was greatly reduced and church construction stopped. When it resumed, it was under the influence of new religious attitudes; the Lutheran religion was introduced into Norway in 1536. The stave churches were considered too small, too dark and too cold. Many were replaced or remodeled; others were not properly maintained and were destroyed by decay. By 1800 there were only about 100 left.

In 1814 Norway, which had been ruled by Denmark for some 400 years, became an independent nation. Independence was accompanied by a revival of national pride and interest in religion. In 1851 a law was passed requiring that all churches be able to hold 60 percent of the congregation, and as a result many older churches were razed. At the same time there was a countervailing interest in the country's cultural heritage, and a few Norwegians began to try to save the surviving stave churches.

One of the leading rescuers was the painter Johan Christian Dahl. When the Vang church in the district of Valdres was about to be torn down, Dahl made a valiant effort to raise the funds needed to buy it. Failing to do so, he persuaded King Frederick William IV of Prussia to buy the church, which was dismantled and reconstructed in a small village in Silesia (now part of Poland). There the church still stands. The episode aroused much interest among the Norwegian people, and in 1844 the Society for the Preservation of Ancient Monuments in Norway was founded to preserve the remaining stave churches. There are currently 29 of them left, many heavily modified or moved to a site other than the original one.

Among the stave churches that stand unmodified and on their original site, however, are two of the most beautiful: the church at Urnes, the one that may represent the genesis of the basilican style, and the church at Borgund. Their survival, along with the results of our survey, shows that the stave churches were well designed and well constructed. They provide confirmation of the fact that it is possible to build wood structures that will last indefinitely if certain conditions are satisfied. The wood must be carefully selected and cured, meticulous design practices must be applied, attention must be given to details that forestall decay, construction methods must be of high quality and the structure must be maintained continuously to minimize deterioration. The evidence shows that if these conditions are met, wood buildings can be permanent.



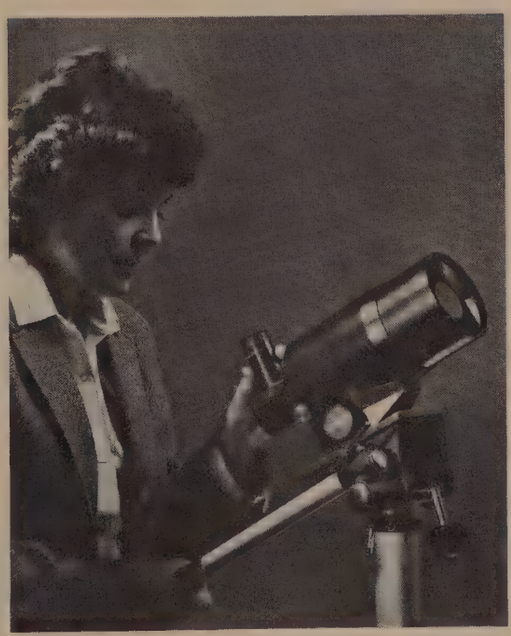
An example of the interesting things that might turn up unexpectedly while you are casually sweeping the deep sky is this supernova in NGC 6946, in a portion of a photograph taken by Hubert Entrop on November 4, 1980, with his Questar. This was 7 days after the official discovery of the nova, as described in Sky & Telescope *in the January, 1981, issue.*

Great news from Questar. . . The Wide-Sky Telescopes

The comet is coming. And Questar is ready for it with two brand-new instruments for all dedicated comet watchers and other deep-sky enthusiasts.

Our emphasis in developing these new designs was on low magnification and optimum field of view—magnification low enough and field of view wide enough to sweep the sky for richfield observing without need of finder or equatorial mount.

We also wanted to dispel the popular misconception that short focal lengths (f/3 to f/5) are essential for effective deep-sky observing and optimum image brightness. Image brightness is a function of aperture and magnification, and has nothing to do with focal length. We continue to use long focal ratios, even in our new wide-sky models. This rules out the comatic aberrations of short-focal-length reflectors and the spurious chromatic effects of short-focal-length refractors, as well as the below-par imagery of short-focal-length eyepieces. Imagine, then, the pleasure of having that familiar sharp Questar® resolution, with no aberrations, available in a telescope with a field of view of 3°.



Specifications of Questar Wide-Sky Telescopes

Questar Wide-Sky 3½ (shown left): aperture 89mm; focal length 700mm; focal ratio f/7.8; magnification 22 × with 32mm wide-angle eyepiece; field of view 3°; faintest visible star 12th magnitude (visual), 12.8 with special coatings.

Questar Wide-Sky 7: aperture 178mm; focal length 2400mm; focal ratio f/13.5; magnification 44 × with 55mm wide-angle eyepiece, 75 × with 32mm wide-angle eyepiece, field of view 1.15° with 55mm eyepiece, 1° with 32mm wide-angle eyepiece; faintest visible star 13th magnitude (visual) with standard coatings, 14th magnitude (visual) with special coatings.

Both instruments can be equipped with various accessories to provide for tracking and photography.

Special uses for the Wide-Sky Questars are comet seeking and general sky scanning, monitoring telescopic meteors, observing the occultation of stars and planets by the Moon, observing lunar and solar eclipses, observing the deep-sky phenomena — nebulae, star clusters and galaxies — and variable star observing.

Photographic accessories are available for use in these applications, in which the Questar becomes an astrographic camera. All Questars, of course, can be used with a video monitor.

Let us send you our price list.

©1983 Questar Corporation

QUESTAR

Box C, Dept. 206, New Hope, PA 18938
(215) 862-5277

QUESTAR, THE WORLD'S FINEST, MOST VERSATILE TELESCOPE, IS DESCRIBED IN OUR BOOKLET IN COLOR, WITH PHOTOGRAPHS BY QUESTAR OWNERS, SEND $2 TO COVER MAILING COSTS ON THIS CONTINENT, BY AIR TO SO. AMERICA, $3.50; EUROPE AND NO. AFRICA, $4; ELSEWHERE, $4.50.

Digital Typography

Most type is now produced not by casting metal or by photography but by computer. The digital typesetter can create new letterforms with the flexibility of a scribe at up to 15,000 characters per second

by Charles Bigelow and Donald Day

The type on this page of *Scientific American* is set by a machine whose operation is digital. It is formed on the screen of a cathode-ray tube by an electron beam moving in a vertical pattern consisting of 800 lines to the horizontal inch; the motion of the beam is fast enough to generate a page of text in about 15 seconds. The resolution of the screen is much higher than that of a standard television receiver, but in other respects the two devices operate in the same way. The tiny individual picture elements—pixels—along each scan line can be made to fluoresce or remain dark by turning the beam on or off as it scans. The resulting pattern of discrete vertical strokes, which would be visible at moderate magnification, is perceived by the unaided eye as a page of smooth letterforms.

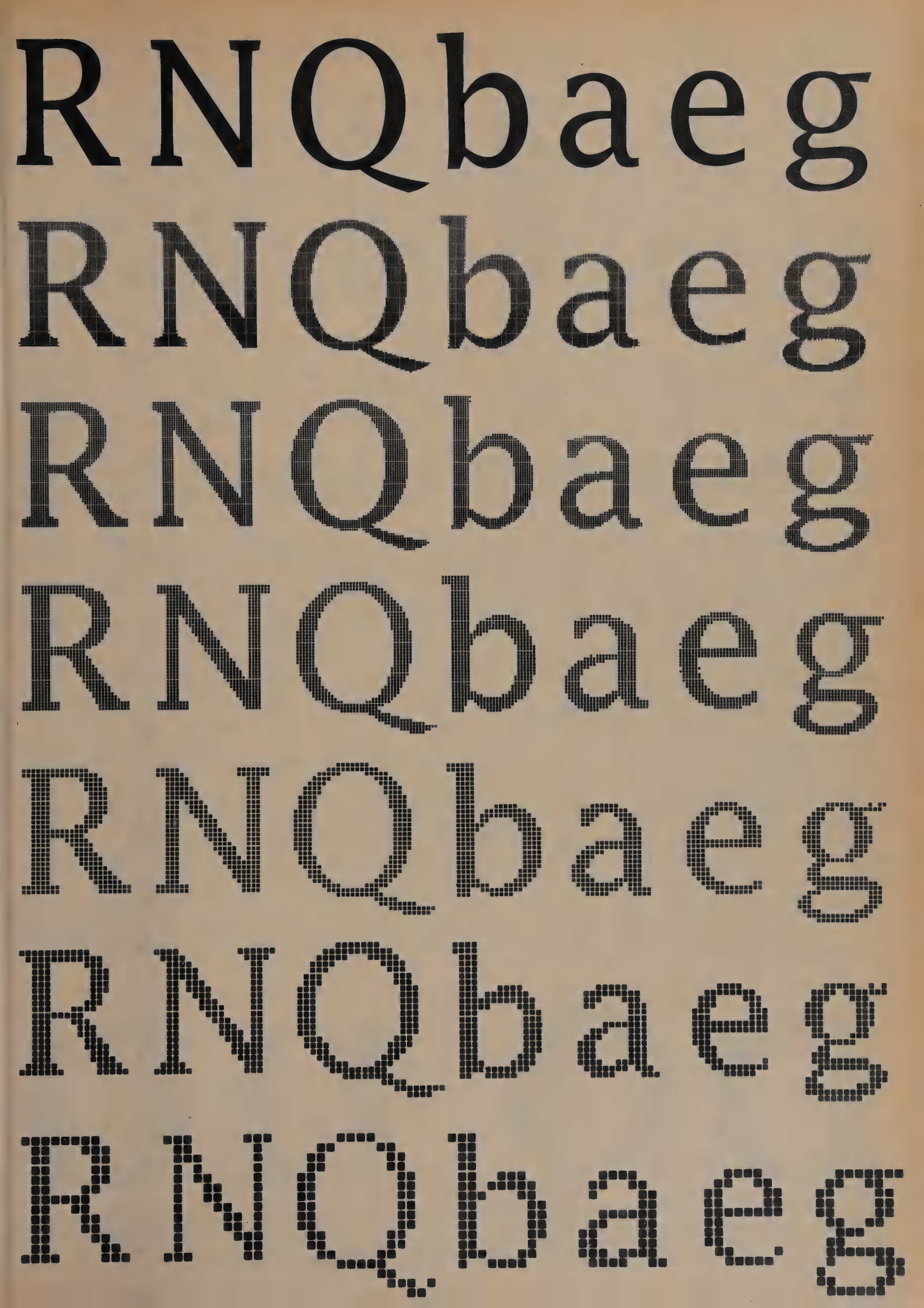
Although a digital computer is needed to control the on-off pattern of the electron beam, the type itself is digital because it is made up of discrete elements. These elements can be line strokes, pixels, colors, shades of gray or any other graphic unit from which a letterform can be constructed. Hence digital typography is not new: mosaic tiles, embroidered samplers and arrays of lights on theater marquees have long represented alphabetic characters as relatively coarse discrete arrays. These digital letterforms, however, are typographic curiosities, far from the mainstream of traditional type design and composition. The traditional letter is not digital but analogue; its final form varies smoothly with the continuous variation of some process employed in its creation, such as the pressure of a brush on paper or the contour of a punch that stamps the matrix or mold used to cast type in metal. With the development of the computer and digital electronics, typography in the past 15 years has been seeing a wholesale replacement of analogue text by digital text, which may rival the shift in the Renaissance from script to print.

It is estimated that each day in the U.S. about 10^{14} letters are reproduced; indeed, letterforms make up much of the visual texture of civilized life. Although some letters are now destined primarily to be "read," or decoded, by a machine, all letters must be read by people if the communication channel between writer and reader is to remain intact. Reading skills, however, are difficult and costly to acquire. In order to protect this educational investment the letters one reads as an adult must not be noticeably different from the letters one learned as a child, and the letterforms read by the current generation must not be significantly different from the ones created by previous generations. The transition to digital typography therefore presents a subtle question with far-reaching implications: How is it possible to take the fullest advantage of digital technology and still ensure that digital letterforms retain the quality of the traditional letters, whose beauty and legibility have contributed profoundly to literacy in our culture?

The advantages of digital typography are substantial: once letterforms are represented as discrete elements they can be efficiently encoded as discrete and distinguishable physical properties in any convenient medium, processed as bits of information by a computer, transmitted over great distances as pulses of current and decoded to reconstitute the letterforms for the person receiving the message. Indeed, once type is digitized it is effectively encoded in the binary language of the computer, and so the size, shape and subtler characteristics of letters can be readily modified by a computer program. Since some varieties of digital type can be read by machines, the semantic content of the information represented by the letters can be manipulated by a computer as well as the letterforms. Unlike analogue information, digital information is highly resistant to noise or degradation introduced during the transmission of a signal. The digital receiving device need distinguish between only two states of the signal (say on or off) in order to decode and recover the information originally transmitted.

Digital typography can be adapted to a wide variety of output devices. The composition of most daily newspapers and national magazines in the U.S. is done by cathode-ray-tube typesetters. A new generation of high-resolution laser printers has now been introduced; a highly collimated laser beam replaces the electron beam as a writing instrument, and it can either expose a typographic image on a printing plate directly from digital information or make an intermediate photographic exposure on paper or film. Lower-resolution laser printers, which are also called text setters, are employed as output devices in data processing and in the publication of forms and documents in small batch quantities. In such devices the laser beam writes out the image of the text by setting up a pattern of electrostatic charge on a belt or drum. Fine pigmented particles called toner particles are attracted to the belt or drum where the charge is created, are transferred to paper like ink from a press and are then

SAMPLES OF DIGITAL LETTERFORMS, shown at decreasing resolution from the second row to the bottom row of the illustration, are each based on the designs in the top row by Kris Holmes and one of the authors (Bigelow). The letters in each row are formed by superposing a square grid of increasing coarseness on the model letterforms at the top. If some part of the model letter coincides with the center of a square on the grid, the entire square is blackened; otherwise the square is left white. Because digital type is constructed out of discrete elements, it is ideally suited for storage, transfer and manipulation by digital electronics systems and computers. Nevertheless, it imposes a new set of constraints on typographic design: the basic letterforms must be "well tempered," that is, they must retain maximum legibility across a variety of digital displays, and so the designer must take into account many versions of the image simultaneously. The digital letters were generated by Autologic, Inc., of Newbury Park, Calif.


RNQbaeg
RNQbaeg
RNQbaeg
RNQbaeg
RNQbaeg
RNQbaeg
RNQbaeg



DUCTAL LETTERS are handwritten letters whose basic topology is the result of a smooth series of movements of the writing tool in the plane of the writing surface. The path of the tool is called ductus. The ductus of each letter is shown in color; the broken lines indicate the part of the ductus that is invisible in the final form of the letter, namely the path of the tool when it is not in contact with the writing sur-

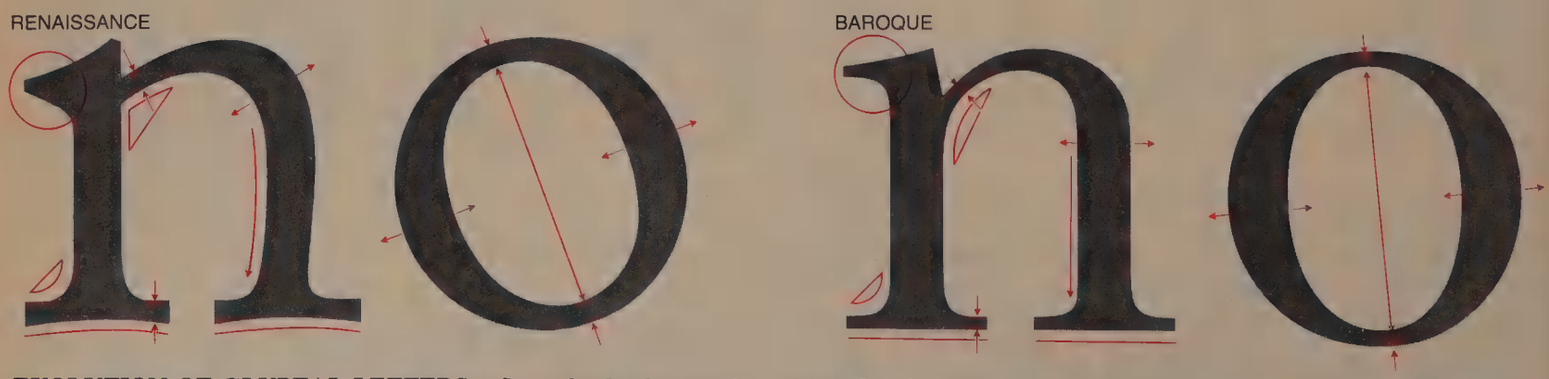
ironed or fused onto the paper by heat.

There are also many devices in service whose output resolution is quite low, that is, far fewer pixels are available to approximate the form of the letter than are required for finely rendered alphabetic details. For example, the printer that is often attached to both large and personal computers is made up of a column of fine wires. The tips of the wires strike a ribbon impregnated with ink and impose a pattern of ink dots in a vertical column on the paper. As the column of wires moves across the page the tips of different wires in the column are actuated by solenoid magnets, and so the changing pattern of vertical dots on the page generates digital text. Ink-jet printers have also been developed that create letter images by directing the flight of electrically charged ink droplets toward the paper. The trajectory of the droplets is controlled by charging them electrically and then passing them through an electrostatic field [see "Ink-Jet Printing," by Larry Kuhn and Robert A. Myers; SCIENTIFIC AMERICAN, April, 1979]. Soft-copy, or transient, images of digital type on cathode-ray tubes, liquid-crystal displays and light-emitting diodes have become increasingly important as the number of personal computer work stations and word-processing devices has grown.

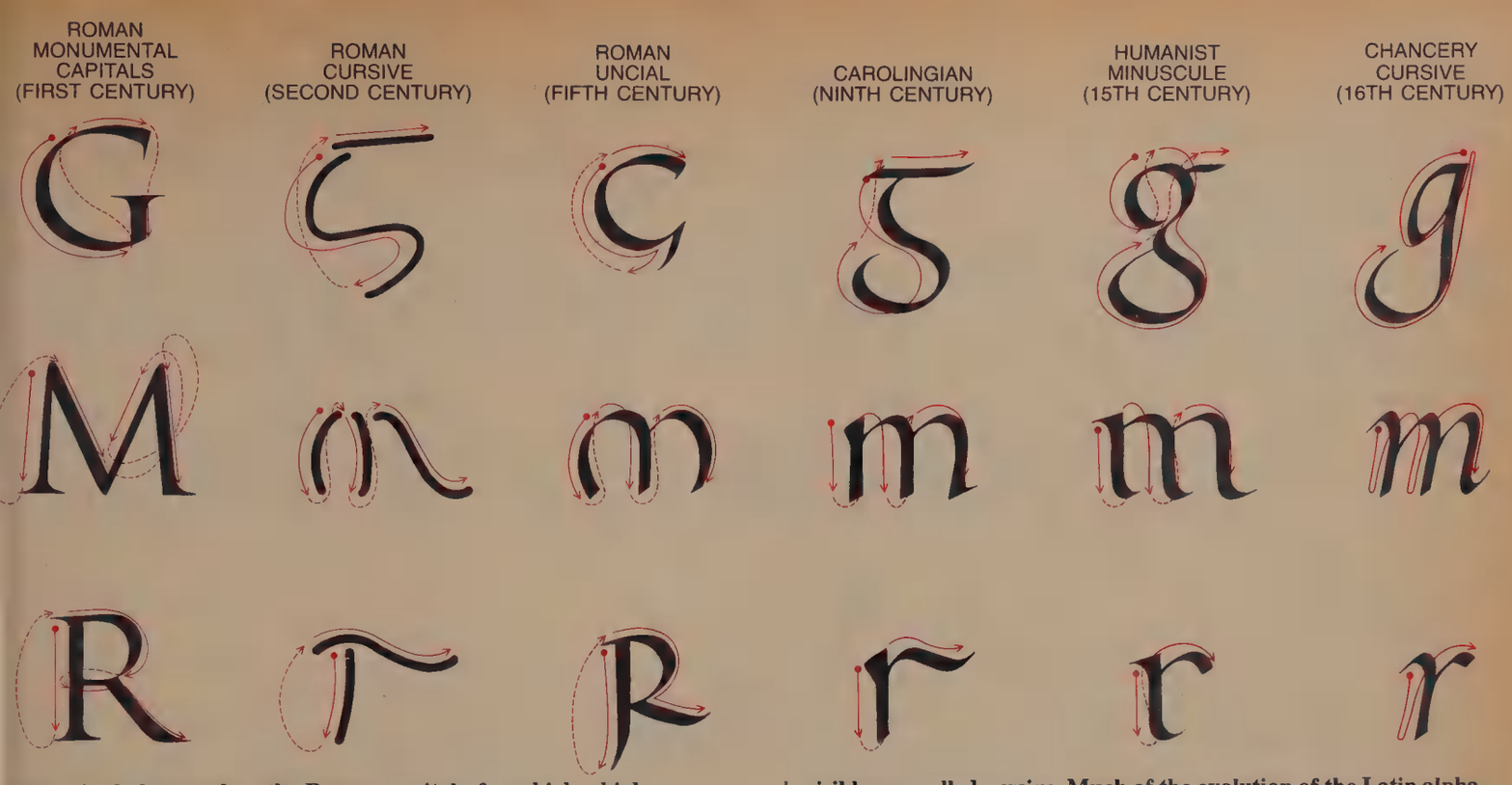
The capacity of all such devices to render typographic characters is best estimated by the number of pixels along the side of a square called an em square, which is equal to the printer's point size of the type in a given font. (There are approximately 72 printer's points to the inch.) The side of an em square, which is also called an em, is slightly more than the distance from the top of the ascender on a letter such as the lowercase *h* to the bottom of the descender on a letter such as the lowercase *y*. For the standard text size called pica, or 12-point type, most high-resolution digital typesetters have a resolution of from 100 to 300 lines per em; electrostatic text setters have a resolution of from 33 to 80 lines per em, and low-resolution digital output devices can vary from 10 to 30 lines per em.

The speed, versatility and low cost of digital typography have made its proliferation irresistible. Moreover, if the letterform is mapped onto a raster, or grid, of digital pixels that is sufficiently fine, the differences between the original letter and the digital letter are virtually indistinguishable. In principle it seems possible, given a fine enough digital raster, to imitate any of the traditional printed or written letterforms, no matter how refined. In practice, however, digital typography, like any other technological innovation, carries with it special problems and a new set of selective pressures.

For example, not all digital typography can begin at high resolution. There are technical limitations on the size of the dot that can currently be reproduced



EVOLUTION OF GLYPTAL LETTERS reflects the fundamental changes in the technology of letter production brought about by the invention of printing from movable type. The letter image was no longer the result of a series of strokes written "on the fly" by a scribe; instead each letter was painstakingly engraved as a single master copy onto the face of a steel punch. Early glyptal letters such as the Renaissance forms imitated scribal forms such as the Humanist minuscule, and many details (*color*) simulate the effects of a broad-edged



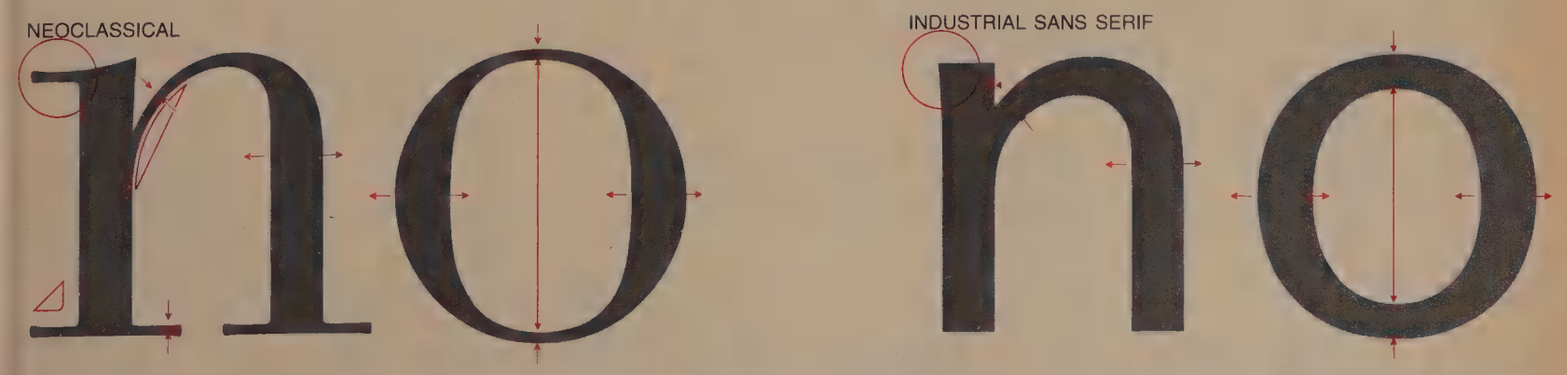
face. Alphabets such as the Roman capitals, for which a high proportion of the ductus is invisible, are called formal, whereas alphabets such as Chancery cursive, for which a high proportion of the ductus is visible, are called cursive. Much of the evolution of the Latin alphabet is the result of the interplay of the opposing tendencies between formal and cursive writing. The calligraphy was done by Holmes.

by methods such as ink-jet printing, and the attainable resolutions are too coarse to escape visual detection. Moreover, the speed and cost advantages of digital typography are reduced as the number of digital bits needed to generate the letter is increased. The number of pixels per letter increases as the square of the linear resolution of the printing device: doubling the linear resolution of the device implies a fourfold increase in the amount of information, or number of bits, that must be transmitted and processed. Although there are computational methods for compressing the data in the bit map of a letterform, the general relation between cost and resolution remains valid.

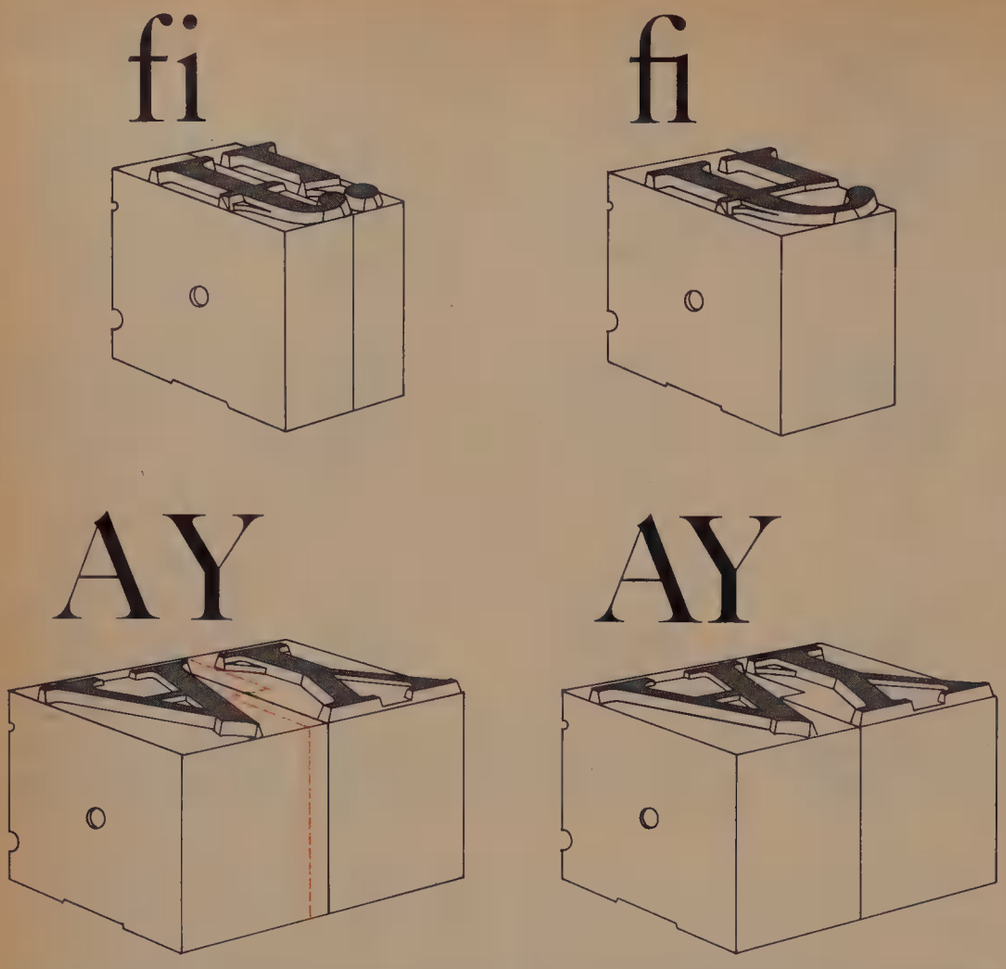
Perhaps the most challenging problem that confronts the digital-type designer is to make effective use of the enormous flexibility inherent in digital technology. For example, the text of a document could first be written at a cathode-ray-tube terminal with a resolution of 10 lines per em for pica type. A proof of the same text could then be pulled from a wire-matrix printer with a resolution of 20 lines per em, circulated for correction and commentary as the output of a laser text setter with a resolution of 50 lines per em and finally set for publication on a cathode-ray-tube typesetter with a resolution of 200 or more lines per em. Similarly, when digital letters must be represented at different sizes on a machine with a fixed raster, a different bit map is required for each letter size. It is obviously desirable in such circumstances for the type designer to create a single kernel, or underlying, letterform that can generate all the forms in which the letter occurs on different machines and in various sizes.

In order to appreciate the magnitude of the problem, consider the variability of letterforms that is reflected in a single superfamily of typeface designs. For each modern design one of each of three opposing features must be specified: whether the type is roman or italic, whether it is normal weight or boldface and whether it is serif or sans-serif. (A serif is a short finishing stroke to a major writing stroke, such as the small horizontal line at the bottom of the stem in the letter *T* as it appears on this page; sans-serif type has no serifs.) Taken together, the three features generate eight typeface designs. Furthermore, each type alphabet typically includes characters in 16 different sizes. The total number of glyphs, or individual bit maps, necessary to accommodate a single character for a minimum superfamily of type is therefore 128; the number of glyphs necessary for a complete superfamily, which may include 128 letterforms, is 128^2, or more than 16,000.

Although digital technology imposes a new set of problems on typographic design, such problems are not without precedent; it is instructive to explore the relative stability of the letterforms



pen held at an angle of approximately 30 degrees to the horizontal. In general the contrast between thick and thin elements is relatively low. In the Baroque and Neoclassical versions of the glyptal letter symmetry, harmony of structure and contrast between thick and thin strokes become more important than the trace of the pen. More recent designs either eliminated serifs, as in the industrial sans-serif types, or converted the serifs into independent elements, as in the industrial slab serif (not shown). The drawings were done by Holmes.



METAL TYPE is cast on a body of fixed width, and so the variant forms and spacing for different combinations of letters that are hallmarks of the fine manuscript cannot easily be emulated. In the first decades after the invention of movable type some printers cast hundreds of variant letterforms, thereby nearly canceling the advantage in time and economy afforded by the new technology. Later forms were well-tempered designs, cast on a metal body in such a way that their assembly into words would automatically lead to proper spacing. A few combinations, such as *fi,* resisted tempering and were cast on a single body of metal. Other letter combinations such as *AY* could not be properly spaced unless the type body was mortised by hand.

6 POINT

RQEN baegnov

8 POINT

RQEN baegnov

14 POINT

RQEN baegnov

18 POINT

RQEN baegnov

SCALING THE SIZE of type cannot be done proportionally even within a single typeface if optimum legibility is to be retained. In the smaller sizes from six to 12 point the letters must be wider and their stems, hairlines and serifs must be thicker than their counterparts in larger sizes. Moreover, in smaller sizes ascenders such as the stem of the lowercase *b* and descenders such as the tail of the lowercase *g* must be shorter in proportion to the height of the lowercase *n;* the counter, or interior space of a letterform, must be more open and the space between letters must be greater. The type font in the illustration is Times Roman; it has been scaled photographically to the same height of the letter *n* so that differences in design can be compared.

through typographic history and their adaptations under the pressures imposed by previous technological shifts. Two stages in the evolution of type design can be recognized following the introduction of each new technology. First, there is a period of imitation, in which the outstanding letterforms of the previous typographic generation serve as models for the new designs. Second, as designers grow more confident and familiar with the new medium, innovative designs emerge that are not merely imitative but exploit the strengths and explore the limitations of the medium. Since type is ultimately intended for the reader, however, the technology of type production is not the only influence on the final letterform; typographic design remains an art, and the successful design subtly reflects the tension between imitation and innovation.

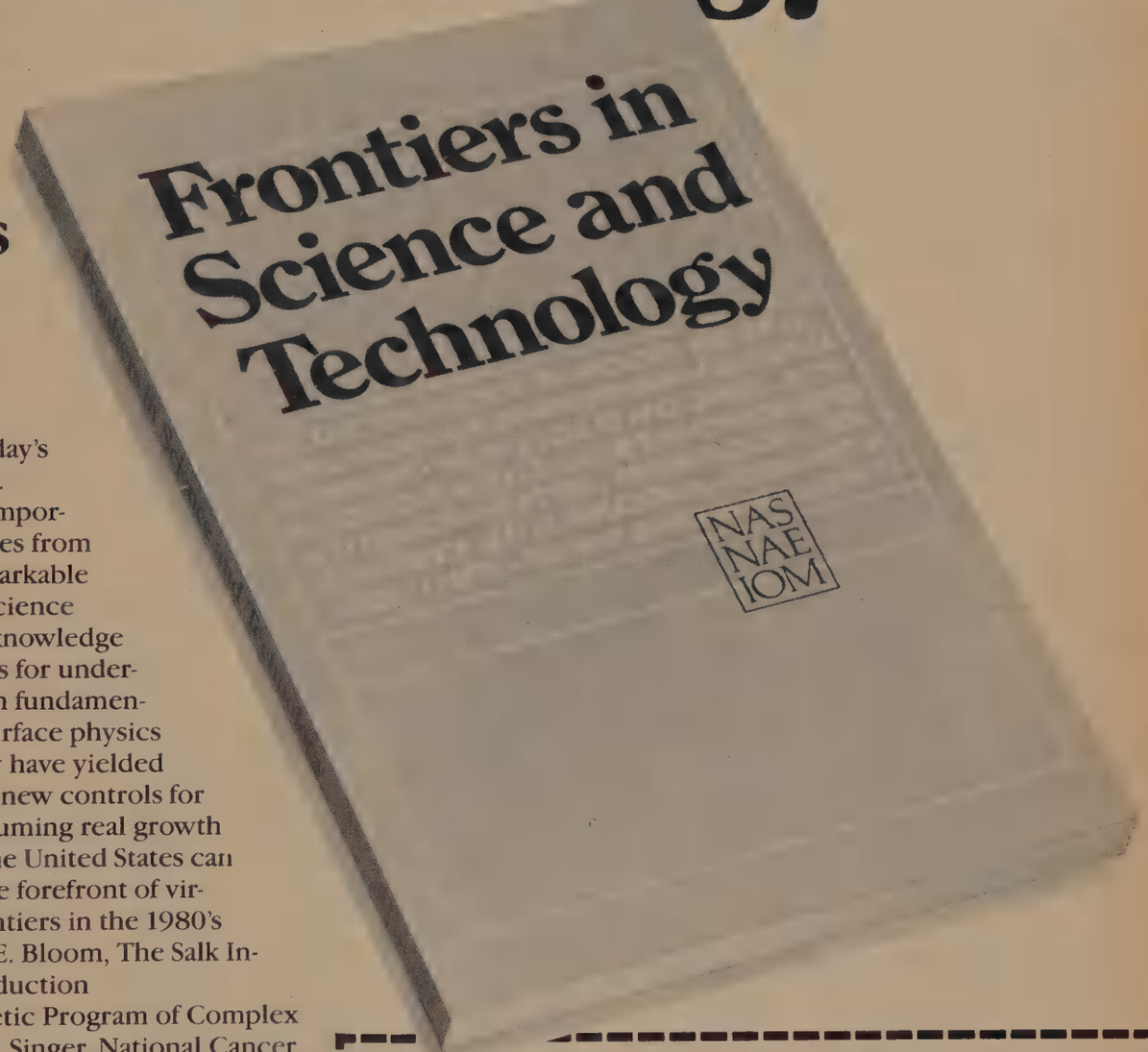
The shapes of letters have persisted longer than any other artifacts in common use. Letter designs are still in service that are more than 2,000 years old, and many common typefaces are replicas of designs popular in the 15th and 16th centuries. The text type in this magazine, for example, was originally designed by the British typographers Stanley Morison and Victor Lardent for *The Times* of London in 1931 and is called Times Roman. It is based on French and Flemish types designed in about 1570. The basic forms of our lowercase alphabet were established in the eighth century in the monasteries and chancelleries of Charlemagne's Frankish empire. The forms of our capital letters are substantially the same as those inscribed by the Romans in the reign of Augustus Caesar. Fully half of our capital letterforms are structurally unaltered from the inscriptional forms used in Periclean Athens in the fifth century B.C.

In the scribal era the contours of the letterform were created by a continuous sequence of movements of a brush or a pen as it moved across the plane of the writing surface. During the smooth sequence of movements the writing tool was pressed against the surface or lifted from it, thereby generating the strokes that make up the letter. The sequential pattern of movements is called ductus, and it defines a characteristic topological structure for each letter of the alphabet. The evolution of writing was impelled by changes in ductus [*see top illustration on preceding two pages*] and by variations in the shape and flexibility of the writing tool. The tool was responsible for the contrast between thick and thin strokes. When the tool was an edged pen, the contrast of the letter was determined by the angle between the edge and the direction in which the pen was moving in the plane of the writing surface. When the tool was a brush or a flexible pointed pen, the contrast in the letter was determined by variations in

An incisive look into our scientific and technological future

Frontiers in Science and Technology

The National Academy of Sciences



The third in a series of reports to the American people, this landmark study examines science and technology's impact on today's most significant issues.

"Perhaps the most important theme that emerges from this volume is the remarkable fertility of American science and technology. New knowledge has provided new ways for understanding diseases. Such fundamental research areas as surface physics and turbulence theory have yielded new technologies and new controls for ancient problems. Assuming real growth in research support, the United States can expect to remain at the forefront of virtually all scientific frontiers in the 1980's and beyond."—Floyd E. Bloom, The Salk Institute, from the Introduction

CONTENTS The Genetic Program of Complex Organisms—Maxine F. Singer, National Cancer Institute. The Molecular and Genetic Technology of Plants—Joseph E. Varner, Washington University. Cell Receptors for Hormones and Neurotransmitters—H. Guy Williams-Ashman, University of Chicago. Psychobiology—Steven Hillyard, University of California, San Diego. Surface Science and Its Applications—Homer D. Hagstrum, Bell Laboratories. Turbulence in Fluids—Willem V. R. Malkus—Massachusetts Institute of Technology. Lasers—C. Kumar N. Patel, Bell Laboratories. The Next Generation of Robots—Jacob T. Schwartz, New York University, and David Grossman, IBM Watson Research Center.

Paperbound. 228 Pages. ISBN 0-7167-1517-1 $14.95

Available at fine bookstores, or use the order form at right.

Yes, please send me _____ copy (copies) of FRONTIERS IN SCIENCE AND TECHNOLOGY at $14.95 each.

_____ I enclose a check for $14.95 for each copy ordered, plus $1.50 to cover postage and handling. (New York, California, and Utah residents please add appropriate sales tax.)

_____ Please charge my ☐ MasterCard ☐ Visa

Account # _____ Expiration Date _____

Signature _____

(All credit card orders must be signed.)

Name _____

Address _____

City _____ State _____ Zip _____

W. H. Freeman and Company
41 Madison Avenue
New York, NY 10010

ISBN: 0-7167-1517-1

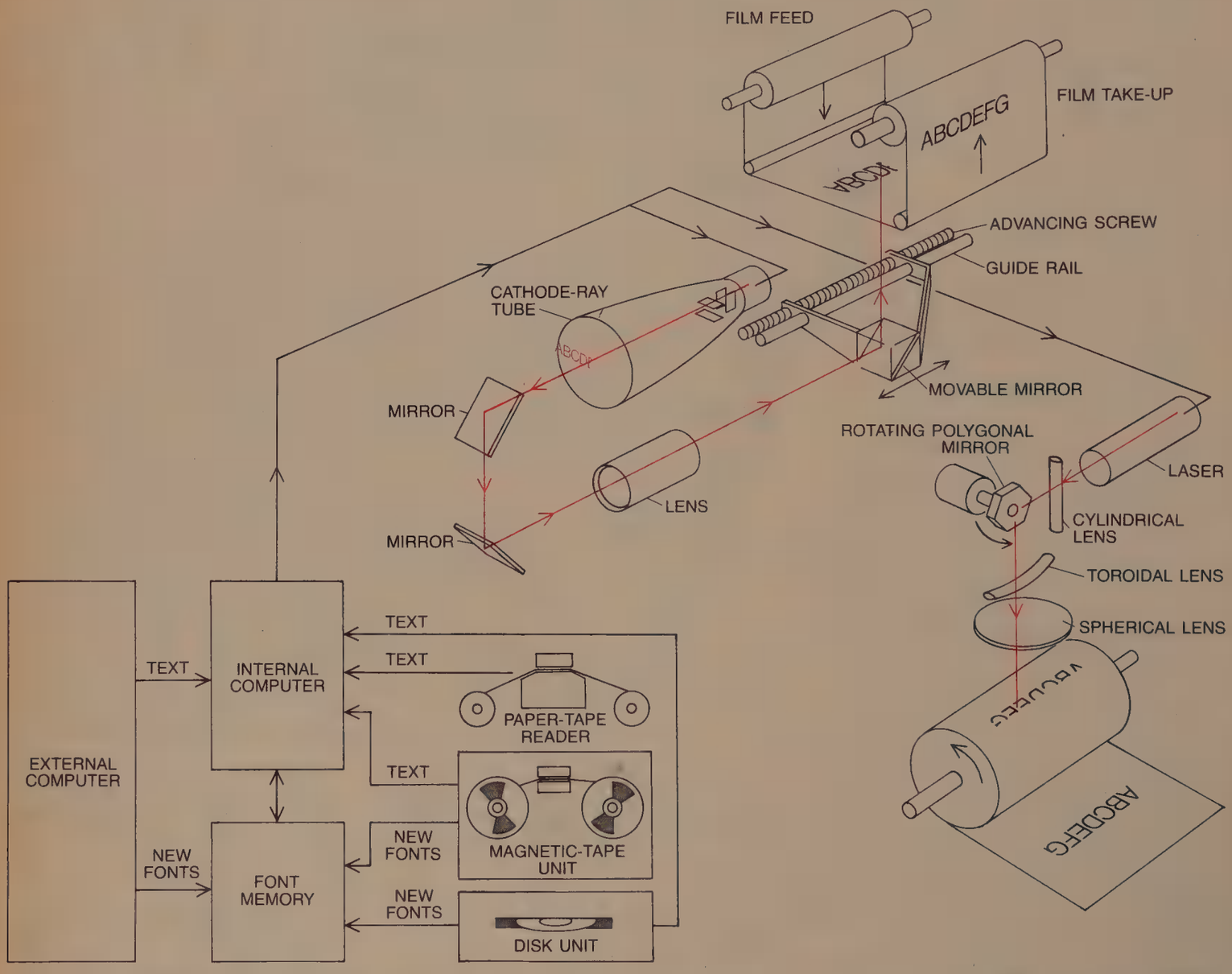
SA

pressure as the tool described its ductus.

An important advantage of scribal lettering is the immediate feedback between the final form of the letter and the scribe. The accomplished scribe is expert in making minor changes in the form of each letter in order to harmonize the letter with the letters adjacent to it. Moreover, to justify lines of text the scribe could insert alternative or abbreviated forms of letters and words. The great disadvantage of scribal practice, of course, is that the design of each letter must be executed in "real time": any letter that is to be read must be generated from scratch each time by hand.

When movable type was invented in Europe, it was at first modeled on scribal forms current in the 15th century, even though it was created by a radically different technology. An image of the letterform was engraved in relief onto the face of a steel punch. The punch was then hardened and struck into a blank of copper to form an intaglio, or recessed, matrix. The matrix was placed in an adjustable mold; when a molten alloy of lead, tin, antimony and copper was poured into the mold and the matrix, the design appeared in relief once again on the face of the cooled alloy. The face of the type was then inked and impressed directly onto paper. The letterform was thus created by a glyptal, or sculptural, process. The letter design was freed from the limitations of real-time execution but was constrained by the need to place it on a rigid rectangular solid. In order to imitate the scribal variations in letterform for different letter combinations a few early hot-metal typographers maintained hundreds of variant letterforms; such forms were so expensive to cast as well as to compose, however, that they nearly negated the economic advantages of movable type.

The evolution of glyptal typeface contours was guided by conceptual and perceptual forces instead of by the needs of rapid handwriting. Greater attention was given to the shape of the spaces within letters and between them. Proportion, width, weight and construction were altered independently of the underlying topology of the letter rather than being partially determined by it, as they were in the ductal letter. The designers of glyptal letters were obliged to make a different engraving for each let-



DIGITAL TYPESETTING MACHINES can store enormous numbers of type fonts for almost instant retrieval and can generate characters in almost any size at rates of up to 15,000 characters per second. The data are entered into the machine by means of punch-coded paper tape, by magnetic tape or disk or as a stream of data from an external computer. In the most recent installations the external computer organizes the text into lines, columns and pages, justifies the text if that is wanted and automatically hyphenates words where it is necessary for good spacing. Once the font, size and position of each character on the page have been specified, the data are assembled into a series of electronic pulses by an internal, "slave" computer, which drives a marking engine. In the schematic diagram two kinds of marking engine are shown. (The two devices would not both be incorporated into any real machine.) In the cathode-ray-tube typesetter an electron beam reconstructs each letter as a series of closely spaced vertical lines. The image is projected onto photosensitive film or paper. In the laser printer a laser beam takes the place of the electron beam; the laser scans horizontally across an entire page at once and generates a pattern of charge on a drum. Toner particles are attracted to the charged regions and ironed directly onto paper by heat.

TAKE THE 3-VOLUME HANDBOOK OF ARTIFICIAL INTELLIGENCE (A $142.40 VALUE) FOR ONLY $3.95

when you join the Library of Computer and Information Sciences.

You simply agree to buy 3 more books—at handsome discounts—within the next 12 months.

Just completed, the massive, 3-volume HANDBOOK OF ARTIFICIAL INTELLIGENCE promises to become the standard reference work in the growing AI field.

Conceived and produced by leading scientists and researchers at Stanford University, with contributions from universities and laboratories across the nation, the *Handbook* makes available to scientists, engineers, students, and hobbyists who are encountering AI for the first time the techniques and concepts in this rapidly expanding computer universe.

The 200 articles cover the emerging issues, technical problems, and design strategies which have been developed during the past 25 years of research. The *Handbook* has been written for people with no background in AI; jargon has been eliminated; and, the hierarchical organization of the book allows the reader to delve deeply into a particular subject or browse the articles which serve as overviews of the various subfields.

The 15 chapters (5 per volume) include: the history, goals, and current areas of research activity; the key concept of "search"; research on "natural languages"; the design of programs that understand spoken language; applications-oriented AI research in science, medicine, and education; automatic programming; models of cognition; automatic deduction; vision and learning research; and, planning and problem solving.

The comprehensive HANDBOOK OF ARTIFICIAL INTELLIGENCE answers questions like:

- What is a "heuristic problem-solving program?"
- How do computers understand English?
- Can computer programs outperform human experts?

AND INCLUDES:

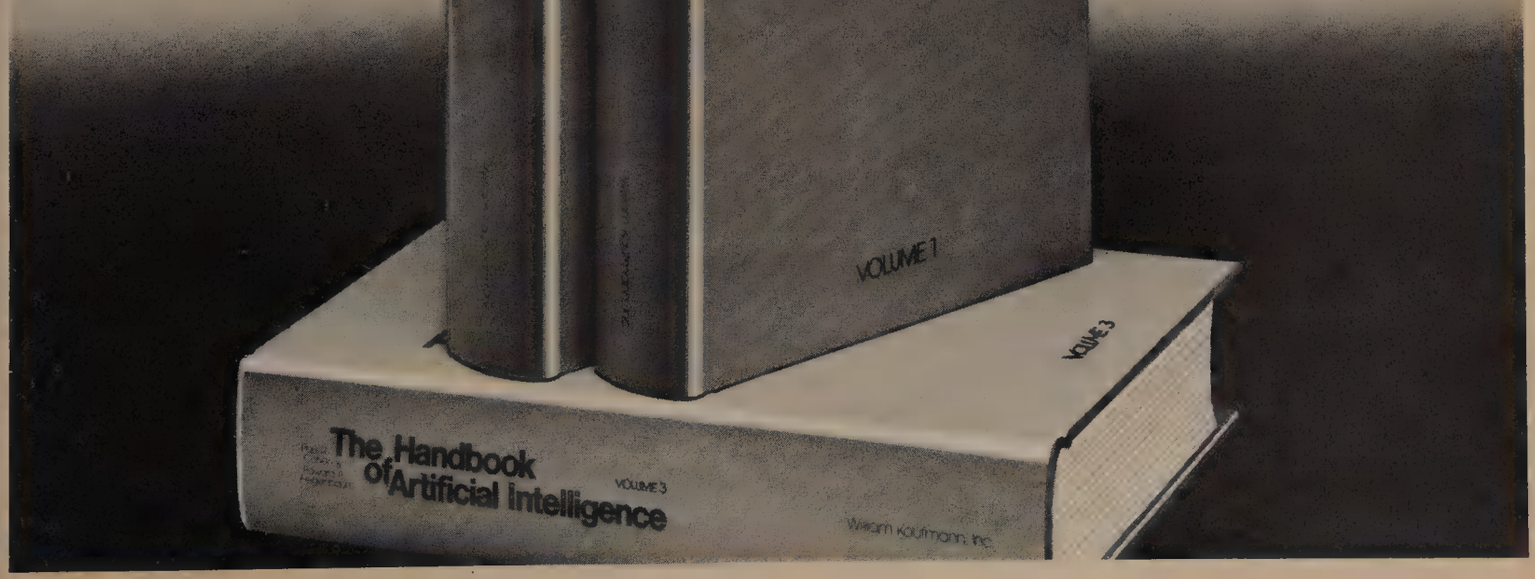
- Over 1,450 pages.
- More than 200 articles in 15 chapters.
- With numerous charts, tables, and schematics.
- Edited by Avron Barr, Paul Cohen and Edward Feigenbaum.

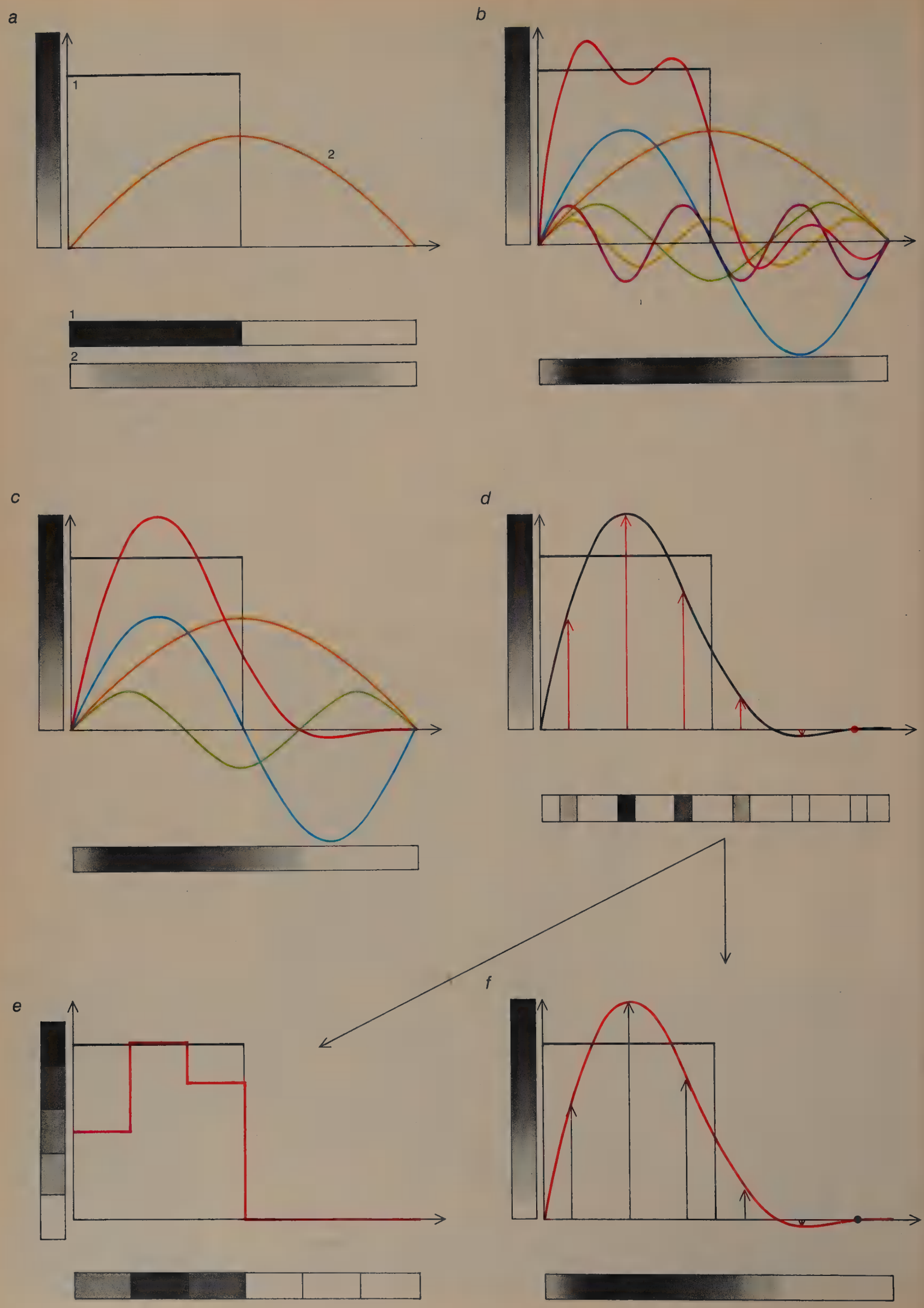
The Library of Computer and Information Sciences is the oldest and largest book club especially designed for the computer professional. In the incredibly fast-moving world of data processing, where up-to-date knowledge is essential, we make it easy for you to keep totally informed on all areas of the information sciences.

Begin enjoying the club's benefits today!

MEMBERSHIP BENEFITS: In addition to getting the 3 volume Handbook of Artificial Intelligence for only $3.95, when you join, you keep saving substantially on the books you buy. Also, you will immediately become eligible to participate in our Bonus Book Plan, with savings of up to 70% off the publishers' prices. At 3-4 week intervals (16 times per year) you will receive the Book Club News, describing the coming Main Selection and Alternate Selections, together with a dated reply card. If you want the Main Selection, do nothing and it will be sent to you automatically. If you prefer another selection, or no book at all, simply indicate your choice on the card, and return it by the date specified. You will have at least 10 days to decide. If, because of late mail delivery of the News, you should receive a book you do not want, we guarantee return postage.

If the reply card has been removed, please write to:
The Library of Computer and Information Sciences
Dept. 7-BV8, Riverside, N.J. 08075
to obtain membership information and an application.




a
1
2
1
2
b
c
d
e
f

ter size, and so, like their scribal forebears, they could make subtle adjustments in spacing and stroke width for different type sizes in order to accommodate the eye. Alternative letterforms were almost entirely eliminated, and the special demands of the new technology were met by letter designs that were felicitous in almost any combination. Nevertheless, the vestiges of the earlier ductal technology were still evident in type fonts cast in the mid-20th century: certain letter combinations such as *fi, fl, ffi* and *ffl* were cast as a ligature, that is, on a single body of metal. To achieve good spacing for combinations such as *TA* and *AY* the body of the type was often kerned, or mortised, by hand at considerable cost.

With the advent of photocomposition in the mid-1950's, typography was confronted once again by a new set of technological variables. In early phototypesetters a stroboscopic light was flashed through a negative master character mounted on transparent film or glass, and from there through a lens that projected an image of the master character onto photosensitive paper or film. The size of the letter image was controlled by interchanging several lenses or by employing a zoom lens of variable focal length. A printing plate was made from the image after the paper or film was photographically developed. Because the position of the character on the paper was controlled by the lens and by a prism or mirror that moved in fine increments across a line of type, the adjustment of interletter spacing for special letter combinations was much simpler than the kerning of metal type. On the other hand, a significant advantage of hot-metal typography was given up. The type design could not be varied with the size of the type without preparing a new master image. The cost of additional master-image carriers and the inconvenience of interchanging them discouraged their use.

The typographer working within the constraints of each of these technologies faced a set of problems quite similar to those now encountered by the digital type designer. For example, any typographic version of the letterforms familiar to the reading public is a successful communication system to the extent that it is able to balance two opposing characteristics: discriminability and similarity of alphabetic characters. To avoid confusing the reader each letter must be rapidly and unambiguously distinguishable from every other letter. On the other hand, the letterforms must also share many graphic features: if a letterform is too distinctive from the other letters in a font of type, it disturbs the flow of reading. Donald E. Knuth of Stanford University has succinctly stated the goal of text-type design: "A font should be sublime in its appearance but subliminal in its effect."

Throughout the history of writing economic pressures have opposed the pressures for more readable and more beautiful typography. As always, therefore, need must be matched to purpose. Letter quality can be sacrificed on certain documents if the tradeoff is faster, cheaper or more compact reproduction. A bill of sale, for example, can be cheap to prepare and to store because it will be read infrequently by few people; the slow and expensive production of an inscription is justified when it will be read frequently by the multitude. In the time of the Romans the basic alphabet was written rapidly in a semilegible cursive when it was employed in a papyrus document, but it was chiseled in clear imperial capitals on a monument. Today inexpensive dot-matrix printout is less readable than the costly typography in mass-market advertising.

The effects on the reader of complex variations in letter design have not yet been quantified, but the response of the visual system to simple spatial variations in light intensity has been studied for more than two decades. The simplest variation to analyze is a visual analogue of an acoustically pure tone, whose periodic variation of intensity with time can be plotted as a sine or cosine wave. A train of spatial sine or cosine waves can be visualized as a ribbon compressed along its length so that its edge traces a series of ordinary sine or cosine waves. If the top side of the ribbon were shaded in such a way that the crests of the ribbon were black, the troughs remained white and the intermediate sections were various shades of gray, then the blurred, parallel bands, or grating, of light and dark that could be seen from above the ribbon would form a spatial sinusoidal wave train.

A pure musical tone is characterized by its amplitude, or loudness, and by its frequency, or pitch. The amplitude of a spatial sine or cosine wave is the maximum contrast, or deviation from neutral gray, that is found in the lightest or darkest parts of the wave train, and the frequency is the number of variations from light to dark and back again within a given distance. Psychophysicists have measured the ability of the visual system to distinguish sinusoidal bands of various contrasts and frequencies from a uniformly gray field. They have found that sensitivity to spatial variation of light and dark depends on the frequency of the variation; the sensitivity is greatest when the spatial frequency is approximately three cycles per degree of visual angle, and no contrast, no matter how strong, can be perceived under most conditions when the frequency is greater than 60 cycles per degree. (The detection of telephone wires against the sky is one of the relatively unusual circumstances in which spatial frequencies that are probably higher than 60 cycles per degree can be discerned.)

What is the importance of these findings for reading? Although the spatial variation from black to white in letterforms is not sinusoidal, fundamental rhythmic patterns in the letterforms are apparent. For example, as the lowercase letter *n* is scanned from left to right across its middle, the brightness of the image varies fairly smoothly from light to the dark stem of the first vertical stroke, then light again in the counter, or interior of the letterform, then dark on the second vertical stroke and finally light again to the right of the letter. Because reading is often done under rather poor lighting conditions, one might expect that the fundamental frequency of text letterforms has evolved to match the peak contrast sensitivity of the visual system.

Such a match is almost exactly what is found in the typesetting of English. In close reading the image given the most attention is projected onto the fovea, the most sensitive area of the retina. The

SPATIAL-FREQUENCY ANALYSIS can be carried out for any two-dimensional image, such as a letter, just as the graph of almost any mathematical function can be approximated by a sum of sines and cosines. A sharp-edged rectangle, for example, can be represented as a straight line parallel to the horizontal axis of the graph; the density of gray or black in the rectangle corresponds to the height of the line above the horizontal axis. In order to approximate the rectangle a unique set of sine and cosine waves can be superposed. The first sine wave in the approximation is shown in *a*. The spatial wave, or variation from white through shades of gray and back to white, that corresponds to the sine wave is a rather crude approximation to the rectangle. The more sine and cosine components there are in the approximation, however, the better it becomes: the five colored sine waves of various heights and wavelengths are added together along each vertical line to give the smooth red curve in *b*. In order to eliminate high-frequency noise any shape can be electronically filtered to remove components above a certain frequency. In *c* the two highest-frequency components of the red curve in *b* have been eliminated, namely the yellow and purple sine waves, and the result is the red curve in *c*. The filtered curve can be sampled by finding its height, or gray-scale value, at evenly spaced intervals along the horizontal axis (*d*). A digital approximation to the filtered curve is made by assigning a discrete shade of gray across an interval that corresponds to each sample (*e*). It can be mathematically proved that the original filtered curve can be completely reconstructed from the samples by interpolating between the sample points, provided the frequency at which the samples are taken is greater than twice that of the highest-frequency component of the filtered curve (*f*).

fovea subtends an angle of one or two degrees, and at a reading distance of 12 inches the subtended angle corresponds to a linear distance only slightly more than the length of a five-letter word set in 10-point type, the commonest type size. There are on the average about 10 spatial cycles across a five-letter word, and so the spatial frequency of the image received at the fovea is about five to 10 cycles per degree, only slightly higher than the most contrast-sensitive frequency of the visual system. There is good evidence that in fast reading word groups longer than five letters can be read in one eye fixation. The reading image is then partially projected onto the parafovea, the region of the retina surrounding the fovea.

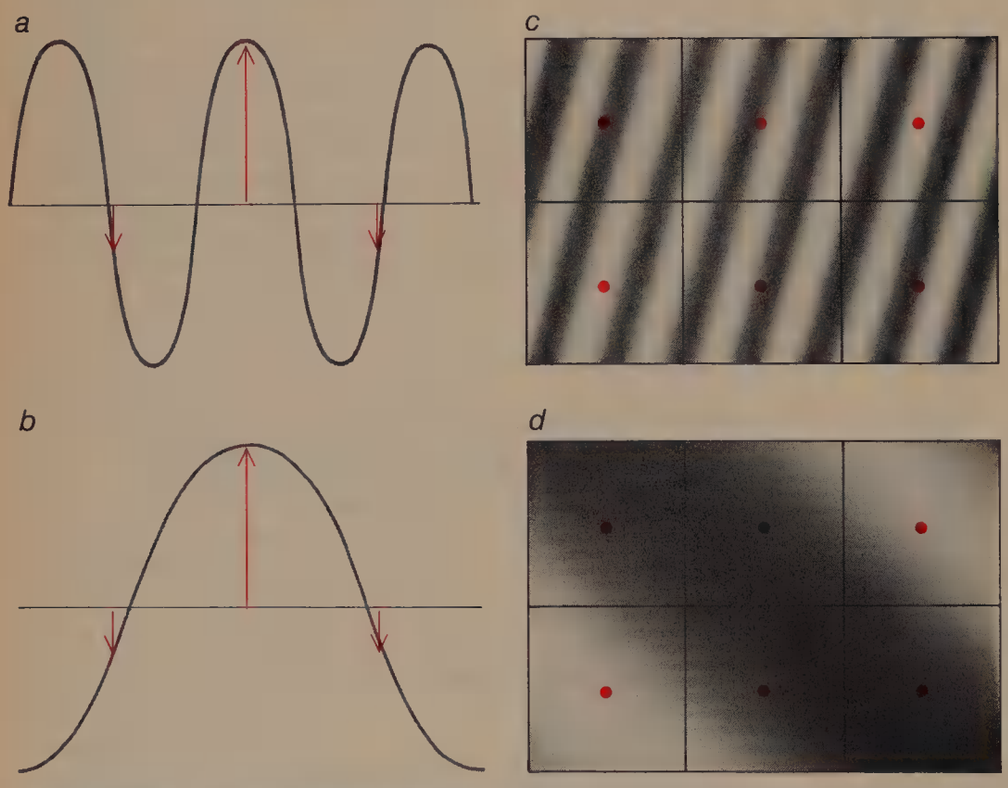
We have tacitly assumed so far that the major patterns of black and white that make up letterforms are sinusoidal spatial waves, but it may not be apparent that the sine-wave model can support a more detailed analysis. It turns out, however, that just as the complex sound of a symphony orchestra can be analyzed as a sum of harmonics, or pure tones of various frequencies and intensities, so can almost any form be analyzed as a combination of many spatial sine and cosine waves. The result follows from a theorem of the French mathematician Jean Baptiste Joseph Fourier. In 1807 Fourier proved that by superposing one-dimensional sine and cosine waves of various phases, amplitudes and frequencies the graph of almost any function can be approximated to any desired degree of accuracy [*see illustration on page 114*].

In two dimensions the spatial sinusoidal components have an additional degree of freedom, namely their orientation with respect to some fixed direction in the plane. Fourier's theorem shows it is possible to reconstruct almost any pattern such as a letterform to any desired accuracy by superposing spatial sine and cosine waves of the proper phase, amplitude, frequency and orientation. In general the high spatial frequencies of a letter image correspond to its edges and to such details as fine serifs and the tapering of the main strokes of the letter. As we have stated, the low spatial frequencies define the fundamental rhythm of the letter, or in other words its overall pattern of light and dark. The spectrum of spatial frequencies required to represent a letter is called the bandwidth of the letter. Based solely on the maximum frequency at which the eye can detect contrast, namely 60 cycles per degree of visual angle, it might seem that completely adequate typographic reproduction could be achieved if all the spatial frequencies of a letterform above 60 cycles per degree were eliminated.

In practice high-quality digital typesetting requires that letterforms include spatial frequencies at least as high as 120 cycles per degree. One reason is that for fine detail such as the hairline at the end of a serif the acuity of the visual system may well exceed 60 cycles per degree. The main reason, however, is the relation between the spatial frequencies that make up the input letter and the process of digitizing the letter: reducing the letterform to an array of discrete units. In order to digitize a letter it must be sampled at various points. For example, if the edges of the letter were perfectly sharp, the letter could be sampled by superposing it on a grid of squares and then noting whether or not the point at the center of each square coincides with some point on the letter. If the center point of a square coincides with a point on the letter, the entire square is shaded black; if the center point does not coincide with any point on the letter, the square is left unshaded.

A perfectly sharp edge would require an infinite number of sines and cosines to represent its full spectral bandwidth. On the other hand, if the bandwidth of a letter is limited in such a way that it includes no sinusoidal component above a certain frequency, its edges must be slightly blurred. The superposition of a finite number of sinusoidal components yields a smooth transition from the black through shades of gray to white. There is nonetheless a mathematical advantage to the band-limited letter: according to results developed by Harry Nyquist at the Bell Telephone Laboratories in 1924 and 1928 and extended by Claude E. Shannon, also of Bell Laboratories, in 1949, if one knows the highest-frequency component of a band-limited signal and if samples are taken with equal spacing at any rate greater than twice the highest-frequency component, the original band-limited letter can be completely reconstructed from the sample points alone. Theoretically, therefore, a letter whose highest frequency is 60 cycles per degree of visual angle can be reconstructed solely from the measured gray values at evenly spaced sample points taken slightly more frequently than 120 times per degree of visual angle.

One difficulty with the theoretical sampling frequency is the small error introduced by the machines that scan



ALIASING is the commonest form of noise, or unwanted signal, found in digital typography. When the height of a wave or the gray-scale values of an image are sampled, information in the original curve or image may be lost. A reconstruction of the wave or the image from the sample points requires that the values of points between the sampled points be interpolated by finding sine and cosine components whose superposition coincides with the values of the sampled points. If the original wave or image is sampled at less than twice the frequency of its highest-frequency component, however, the highest-frequency components found by the reconstruction will have an alias frequency, lower than the highest-frequency component in the original image. In other words, spurious low-frequency components will replace the true high-frequency ones. For example, if a sine wave is sampled only 1.5 times per cycle (*a*), a reconstruction that matches the values of the samples (*colored arrows*) is a sine wave of half the original frequency (*b*). In the two-dimensional spatial wave the samples are gray-scale values taken at the centers of squares (*c*). A reconstruction of the image from the sampled values generates an alias not only of lower frequency than the original waves but also oriented in a different direction (*d*). Quantizing a letter image only compounds the errors introduced by aliasing.

and sample the input letters. Sample points may not be evenly spaced and the measured values of gray may be slightly inaccurate. A more serious problem is that, in practice, sharp band limiting is impossible. Hence in typography of the highest quality the samples are taken at a rate of at least 240 per degree of visual angle, which is theoretically sufficient to reconstruct a letter band-limited at 120 cycles per degree. Such a sampling rate is equivalent to a resolution of 200 lines per em for 12-point type, or 1,200 lines per inch.

If the letter image is sampled too sparsely, that is, at less than twice the frequency of its highest sinusoidal component, the gray-scale values of the samples can be identical with the values that would be given by a lower spatial frequency [*see illustration on opposite page*]. In such undersampling noise is added and information is inevitably lost. Low-frequency spatial waves that do not figure in the original wave spectrum of the letter replace the original high-frequency components; a reconstruction of the original letter from such samples would incorporate the spurious low-frequency components. The phenomenon is called aliasing, and it is the most obvious source of noise or distortion in digital typography. When the letter is converted into an array of pixels from the samples, the aliasing becomes manifest, and the amount of aliasing depends on the coarseness of the sampling.

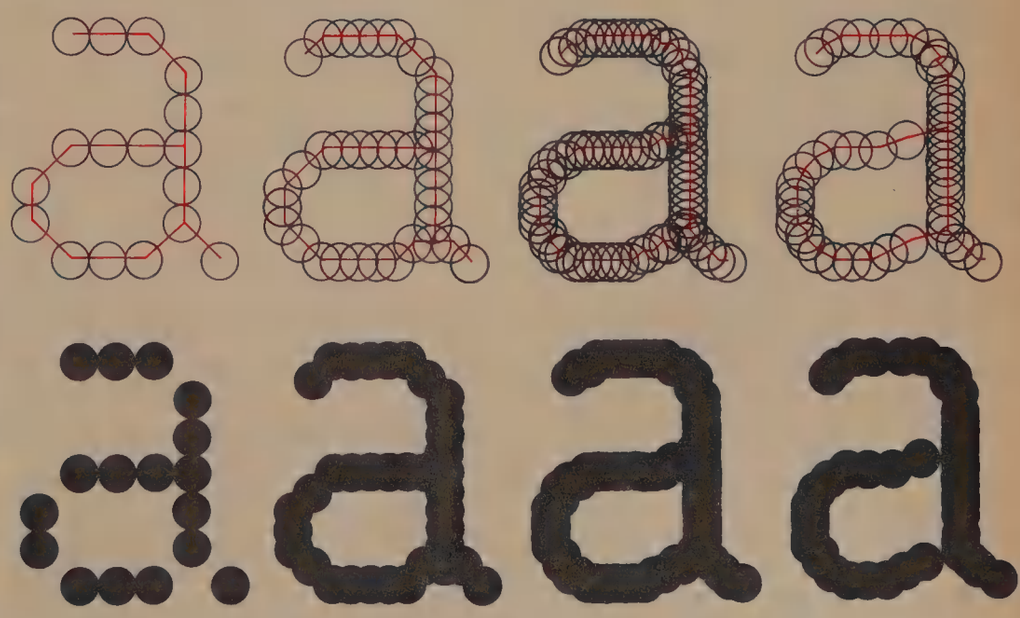
At a high output resolution aliasing is evident only as a slight roughness in the letter contours; at medium resolution the contours become jagged, and at lower resolution the curves become polygonal and the diagonals develop dislocations. At still lower resolution the differences among straight, curved and diagonal elements is obscured and the letters become illegible. Moreover, as the resolution becomes coarser the diversity of possible letter designs is reduced; for very coarse resolutions, such as the five-by-seven or six-by-nine arrays of dots on many cathode-ray-tube terminals and dot-matrix printers, few variant designs are possible.

A less obvious but ultimately more serious consequence of undersampling is the loss of information about the proportional relations reflected in an alphabet design. In a design that has been finely tuned to the characteristics of the visual system there is a subtle interplay of proportion and thickness among the thick elements of a letterform and between thick and thin elements. The ratio of stem height to the height of the lowercase letter *x* and to the average width of a letter is also carefully adjusted. The outcome of such detail in design is barely perceptible in the individual letterform, but in a block of text the design becomes manifest as a visually harmonious pattern of black letterforms and white counterforms and a pleasing level of gray in the text as a whole. At a high digital resolution these proportional variations can be closely approximated, and the typographic texture appears only slightly less refined than it is in analogue typography. As the digital resolution of a letterform is decreased, however, proportional variations are rounded off, typographic elements become homogeneous and the resulting textual pattern seems crude and awkward.

When cost or the technical limits of



QUANTIZATION ERROR results when the shade of gray of a pixel, or minimum discrete picture element, does not match the shade of gray at the sampled point corresponding to the pixel on the original image of a letter. In most digital typography, for example, the pixel must be either black or white, which reflects the naive view that the edges of a master letterform are perfectly sharp. Actually the spatial frequencies too high to be perceived by the visual system should be filtered out before the letter is sampled and digitized. The edges of the filtered letter that results are not sharp; there is instead a continuous tone of gray that makes the transition from black to white. When the letterform is reproduced on a machine of relatively low resolution, such as a cathode-ray-tube terminal, the apparent degradation of the letter caused by the low-resolution sampling can be reduced if the pixels more accurately reproduce the shade of gray measured at each sample point. The pixels that make up the letterform at the left must be either black or white, and the design is almost illegible. The pixels that make up the same letterform can take on one of 16 gray values in the design at the right. By squinting and observing the design at the right from a distance of about 25 feet an ampersand can clearly be perceived. The digital designs were prepared by John E. Warnock of the Xerox Corporation.



SIZE OF A WRITING SPOT in a digital printer such as a dot-matrix printer determines the smallest resolvable element of the letterform, but the position of the spot can still be adjusted to improve the image. The parallel rows of wires on the printer head can be staggered or the head can be made to pass several times across the same character. At the left a lowercase *a* is reproduced without overlap of the writing spot, and the result is a letter of poor quality. The overlap of the writing spot is increased to 50 percent in the second letter from the left and to 75 percent in the third one; the resulting letterforms are more continuous and the edges of the letters are better defined. As the overlap increases, however, the horizontal lines become too thick with respect to the vertical lines and the joins become too blotchy. At the right certain spots have been removed in order to lighten the horizontals and streamline the joins.

the output device require that the letterform be stored or reproduced at less than ideal resolution, there are still several strategies that can be followed to improve the image of the letter. For example, if the shade of gray measured at each sample point were preserved across the entire pixel corresponding to the point, the jagged lines would be much less apparent in the final letter than they are when the pixel must be either black or white. Some devices that have low spatial resolution can reproduce gray-shaded pixels, and so what is called quantization error between the shade of the pixel and the shade of the sample point can be made quite small. The video-terminal screen, for example, can have 256 or more gray levels for each pixel, which can help to compensate for low sampling resolution.

A second strategy for improving the image of a letterform is to store the letter in the memory of a machine in such a way as to take advantage of the overall regularity in its design. Most letters, for example, are simple and connected forms; information about the gray-scale value of the samples inside one of the stems of a letter is redundant because all the samples are black. Hence the space in the memory of a computer that would be necessary to store an arbitrarily complex set of gray-scale values for the pixels making up the stem can be reassigned to store more information about its boundaries. Letters can also be structurally coded as combinations of parts that correspond to the pen or brush strokes in the handwritten forms. More sophisticated coding of the regularity in letterforms requires less storage in computer memory but more effort to decode it later.

One of the most versatile ways to reduce the number of samples that must be stored is to record only the positions of a few selected points along the outline of a letter. The points can then be joined by mathematically constructing lines or curves called splines, which can be computed from the outline points when the letter is retrieved from the computer memory. If only linear splines are employed, the curves of the letter are rendered as polygons; the line segments must therefore be short enough for the polygonization not to be discerned. Curved splines derived from the graphs of quadratic equations such as the circle, the ellipse and the parabola yield a better approximation to the complex outlines drawn by hand than linear splines do. Higher-order splines such as those derived from the graphs of cubic equations can give even closer approximations, but they require more computation to construct.

When the letterform is to be stored as a set of spline knots, or points, the coordinates of the points can be encoded by a device called a digitizer tablet at resolutions as high as 8,000 lines per em, much higher than the resolution of any text-output device. The outline of the character given by the splines can then be employed as a template for generating an array of pixels for any scanning pattern. The stored set of spline knots can also serve as a generic letterform from which a computer can generate a wide variety of particular versions of the letter. For example, higher-order splines can be converted into circular arcs or straight lines on devices that generate a letter outline rather than a bit map. The contours of the letter can be scaled up or down and stretched in any direction. If the position of each spline knot is supplemented by a descriptive label that states its role in the shape of the letter with respect to other spline knots, a typeface can be automatically fitted to a particular raster without the tedious job of turning pixels on or off by hand.

Several computer programs have been written that can manipulate spline-based letterforms. The Ikarus system, developed by Peter Karow of URW Unternehmensberatung in Hamburg, is widely used in the typographic industry. A precise drawing of the outline of a letter on which the spline knots have been marked is placed on an electronic grid, and the positions of the knots are entered into computer memory with a cursor. Letter elements such as stems and serifs are identified as they are entered. The Ikarus program computes the spline outlines, and it can automatically carry out a number of design variations, such as changing the size of a letter on a fixed scanning pattern, varying the thickness of the letter strokes from light or medium to boldface and interpolating between the differing forms of the same letter in two type fonts.

Another spline-based design system called Metafont has been developed by Donald Knuth. The language employed for programming in Metafont is based



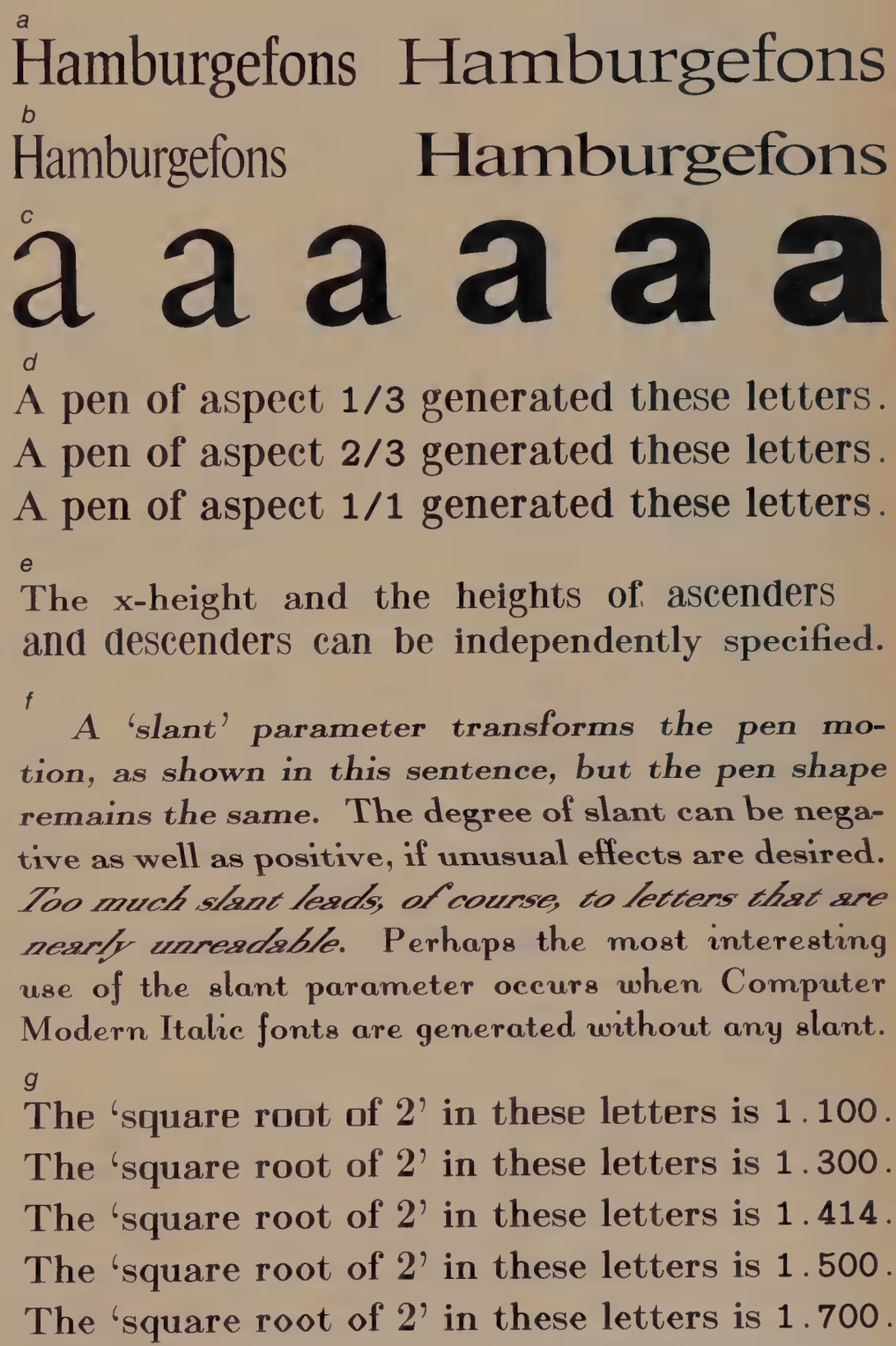
STORAGE SPACE in the memory of a computer necessary to reproduce a letter at a given resolution can be reduced in several ways. One strategy is called run-length encoding, in which only the endpoint positions of each stroke of an electron beam or a laser beam are stored instead of the gray-scale value of each pixel (*left*). Spline encoding can further reduce the memory requirements: the outline of the letterform can be specified at several critical points called spline knots. When the letter is needed, a computer can then interpolate straight lines (*middle*) or straight lines and circular arcs (*right*) between the spline knots in order to generate the endpoints for each stroke of the writing instrument. More elaborate curves such as logarithmic spirals can also be interpolated in order to more closely approximate the original letterform, but such curves trade a reduction in the computer memory for increased calculation "on the fly."

on ductal principles; once the topology of the letterform has been described the final form of the letter is determined by specifying the characteristics of a virtual "pen" that traces the skeleton of the letter. (The virtual pen must not be confused with any real pen; the computer simply represents the final shape of a letter as if it had been drawn with a real pen.) The size of the virtual pen, its angle with respect to the writing surface, the shape of its tip and other variables that can be independently specified generate quite different letter contours from the same skeleton. Other versions of Metafont are under development that describe the letterforms according to glyptal principles as well as ductal ones.

Several other systems have also been realized: FRED, a cubic spline program written by Patrick Baudelaire, and Prepress, a pixel editing program written by Robert F. Sproull, were developed for the Alto computer work station at the Palo Alto Research Center of the Xerox Corporation. A spline system based on spiral curves was developed by Peter Purdy and Ronald McIntosh of Purdy and McIntosh in Watford, England, and the ELF system, based on ductal principles, was written by David Kindersley and by Neil Wiseman of the University of Cambridge. A program and computer work station called the Letter Image Processor has been developed by the Camex Corporation in Boston.

In spite of its almost universal application and widely recognized flexibility, digital typography is only now beginning to move from the imitative to the innovative phase of its development. The sampling of letters and indeed the entire theoretical apparatus on which the sampling is based presuppose an already existing letter design. This model letter is an analogue form, and the success of the digital letter is still judged almost entirely on the degree to which it imitates the ductal or the glyptal letter. Nevertheless, the problems inherent in sampling and digitizing such a letter suggest it would be more productive to design new letterforms directly for digital technology. Moreover, recent advances in the study of vision create an opportunity for the digital-type designer to experiment with ways of adapting typography even more closely to the needs of the visual system. If letter images can be developed that more closely approximate the "language" of visual perception, the speed and efficiency of reading can be enhanced.

Creative design can best be accomplished on a synthetic system in which the type designer can interact rapidly with a computer and immediately see the effects of design changes on a screen, much as the traditional designer can immediately see and correct the mark of a pen or a brush. The system must be precise, high in resolution and capable of instantly reproducing the design in many versions, such as a spline-based outline, a bit map, a gray-scaled pixel array or a simulation of the output of a particular printer. No such machine is yet at hand, but the emerging generation of special integrated circuits for use in the graphic arts may soon make it possible to build one. When such systems become available, there will surely be a flowering of new letterforms as the digital era, like the ductal and glyptal eras before it, enters its creative phase.



VARIATIONS IN THE DESIGN of digital letterforms can be carried out automatically with the aid of several computer programs. In the Ikarus system, developed by Peter Karow of URW Unternehmensberatung in Hamburg, letters can be compressed or expanded without altering the width of their stems (*a*). The program can also correct the design automatically for different printing sizes; the required changes are much more complex than simply stretching the image in one direction (*b*). Smooth interpolations can be generated between a given letter in two different fonts. In *c* the letter at the left is Bembo and the letter at the right is Helvetica Black; the intermediate forms were constructed by the Ikarus program. Another design system called Metafont has been developed by Donald E. Knuth of Stanford University. A program written in the Metafont language alters certain characteristics of the letters by controlling the properties of a virtual pen, by means of which the computer represents the final form of the letter. For example, the shape of the tip of the pen can be controlled; it is an ellipse for which the aspect ratio, or the ratio of the vertical axis to the horizontal axis, must be specified (*d*). Samples of the effects of additional design variables on letterforms are also shown (*e*, *f*, *g*).

THE AMATEUR SCIENTIST

In which simple equations show whether a knot will hold or slip

by Jearl Walker

A hitch is a knot intended to secure a line to another line or to a fixture. Are some hitches securer than others? Is there a limit to the load that can be placed on a hitch before it fails because the free end slips through the knot? Is there some way to alter a normal hitch so that it can withstand a heavier load? In exploring these questions I am guided by the work of Benjamin F. Bayman of the University of Minnesota, who has published an impressive theory about the strength of hitches.

The security of a hitch depends primarily on two features of the knot: how it is wrapped around an object (in what I call wraparounds) and how it passes over itself (wrapovers). Both features provide friction to maintain the hitch when the cord is put under tension. Without friction the cord slips through the knot, untying itself. My discussion is limited to cords of moderate diameter tied around fixed rods of larger diameter. What degree of friction in wraparounds and wrapovers will ensure that a hitch does not fail even when an arbitrarily large load is put on the cord?

I shall disregard here several practical aspects of knots. For example, some knots are better suited to a load that pulls perpendicular to the rod; others are designed for a load that pulls parallel to the rod. Some knots offer a certain advantage in holding for a long time even though they are slipping. Certain types of knots may be better if the load varies periodically in strength. All these factors are of secondary importance here. I shall also disregard any additional strength in a hitch that derives from the way adjacent sections of cord rub against each other (except at a wrapover). The principal sources of friction in a hitch are in the wraparounds and wrapovers.

To understand the role of tension in a cord under load imagine holding up a weight by a cord. Tension means that any small section of the cord is being pulled in opposite directions by forces from the adjacent sections. The tension is uniform along the length of the cord. Hence the force applied at the upper end of the cord must match the weight attached to the lower end if the assembly is to be stationary.

Next imagine holding up the weight by passing the cord over a fixed horizontal rod. The section of the cord between the weight and the rod is under uniform tension but the tension of the section in contact with the rod varies, decreasing from the side with the weight to the other point where the cord leaves the rod. The tension of the section between the rod and your hand is uniform but is lower than it is on the other side of the rod because of friction between the cord and the rod. Because the cord rubs against the rod, you no longer have to apply as large a force to hold the weight in place.

The amount of friction on the cord depends partly on the roughness of the two surfaces in contact and also on how much of the cord touches the rod. You gain friction by wrapping the cord several times around the rod in the maneuver known to sailors and climbers as belaying. Each additional wraparound reduces the force you must apply at your end of the cord.

Friction is also generated in a hitch if the cord passes over itself. When the cord is put under tension, the bottom part of a wrapover is pressed against the rod by the top part. The squeeze opposes the tendency of the cord to slide. The tighter the squeeze, the greater the friction. The hitch holds fast provided the friction is large enough to nullify the force that tends to make the bottom part of the cord slide through the wrapover.

A hitch is designed so that the friction generated within the knot is great enough to hold against the tension generated by a load. Suppose the load increases steadily. Does the tension eventually rise to the point where the free end of the cord begins to slide through the knot, or does the increase in tension merely increase the friction so that the cord cannot slip at all? If the friction increases automatically, the knot is self-tightening and cannot fail under any load. The cord breaks first. Whether the hitch fails or holds depends on its topology and on the friction provided by the wraparounds and wrapovers.

Consider a wraparound that is an integral multiple of a revolution. On one side the tension in the cord is large; call it T_2. The tension is less on the other side; call it T_1. The cord will not slip over the rod if T_2 does not exceed a certain multiple of T_1. The multiple is an exponential function of two quantities: the angle (in radians) through which the cord is wrapped around the rod and the coefficient of friction between the cord and the rod. If the cord makes one revolution around the rod, the multiple is the exponential function of 2π radians multiplied by the coefficient of friction. For simplicity the function can be written as e. The cord will not slip if T_2 does not exceed $e \times T_1$. If T_2 is larger than this amount, the friction between the cord and the rod is overwhelmed and the cord slips.

If the cord is wrapped twice around the rod, the multiple is an exponential function: 4π times the coefficient of friction. A simple way to write the function is e^2. The cord will not slip provided T_2 does not exceed $e^2 \times T_1$. This relation shows the advantage of an additional wraparound. T_2 can now be much larger than T_1 because the additional friction in the extra revolution of wraparound helps to hold the cord in place.

For any integral number of revolutions of the cord around the rod the relation between the tensions on the two sides of the wraparound is similar in form. Let n be the number of revolutions. The cord will not slip provided T_2 does not exceed $e^n \times T_1$.

The other method of gaining friction in a hitch is with a wrapover. Since the top part of a wrapover squeezes the bottom part against the rod, the bottom part resists being pulled through the wrapover to loosen the hitch. Consider the tensions on the two sides of the bottom section in a generalized wrapover. On the side with the large tension (T_2) the force tending to make the bottom part of the cord slide through the wrapover is equal to T_2.

Opposing this tendency is the smaller tension (T_1) on the other side of the wrapover. An additional opposing factor is the friction provided by the squeeze from the wrapover. At most the friction can be a certain multiple (u) of the tension in the top part of the wrapover. The bottom section of cord does not slip through the wrapover if T_2 does not exceed $T_1 + uT$. If T_2 is too large, the friction from the pressure of the top section is not enough to hold the bottom section in place.

The value for u is difficult to calculate for a general situation because it involves the coefficients of friction between the rod and the bottom cord and

Systems Network Architecture

IBM

To: Gina
From: Bill
Subject: IBM Technology

Here's the partial list I promised you of our past and present technological achievements. There are lots of things here that should be of real interest to the scientific, engineering and academic communities. What's your choice for the next topic in this series?

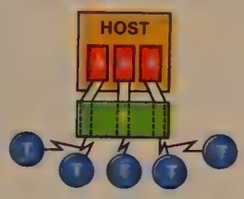
Vacuum tube digital multiplier
IBM 603/604 calculators
Selective Sequence Electronic Calculator (SSEC)
Tape drive vacuum column
Naval Ordnance Research Calculator (NORC)
Input/output channel
IBM 608 transistor calculator
FORTRAN
RAMAC and disks
First automated transistor production
Chain and train printers
Input/Output Control System (IOCS)
STRETCH computer
"Selectric" typewriter
SABRE airline reservation system
Removable disk pack
Virtual machine concept
Hypertape

System/360 compatible family
Operating System/360
Solid Logic Technology
System/360 Model 67/Time-Sharing System
One-transistor memory cell
Cache memory
Relational data base
First all-monolithic main memory
Thin-film recording head
Floppy disk
Tape group code recording
Systems Network Architecture
Federal cryptographic standard
Laser/electrophotographic printer
First 64K-bit chip mass production
First E-beam direct-write chip production
Thermal Conduction Module
288K-bit memory chip
Robotic control language

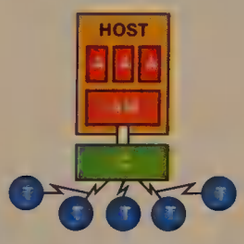
Bill -
SNA is becoming more important every day.
Let's tell that story.
Gina

Figure 1. *EVOLUTION OF SNA NETWORKS*

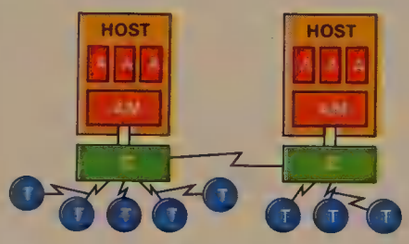
Legend:
A = *application program*
AM = *access method*
C = *communications controller (concentrator/multiplexer/front-end)*
T = *terminal or peripheral processor*
——— *channel connection*
—Z— *communications link*



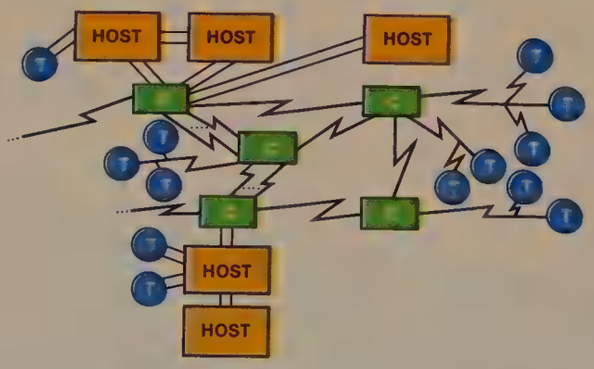
(a) In a typical pre-SNA network, communications links and terminals were dedicated to single uses or applications. All terminals on a link had to connect to the same application program, which included communications software. Usually, changing the terminal or link connections also forced the application programs to be changed.



(b) Early SNA introduced sharing of links among various application programs. A host access method permitted easy access from any terminal to any application program in the host processor. The connections could be readily changed without affecting the application programs.



(c) Subsequently, SNA configurations were enhanced to allow access between host processors for distributed processing and data-base sharing. Moreover, any terminal could access any application program at any host.



(d) Today, SNA networks can be fully meshed configurations. Parallel links between adjacent communications controllers allow increased network availability and traffic balancing. Access from host to host and terminal to host is permitted over multiple routes. The number of different types of network nodes has increased considerably, particularly among terminals and peripheral processors. SNA networks include open interconnection of both IBM and non-IBM nodes.

Advances in computing, processing and communications technologies have prompted increased interconnection of terminals, processors and communications facilities.

These various devices have been linked into networks for distributed access to processing and data-base resources.

A variety of networking applications has been developed for airline reservations, banking, store checkout, process control, remote job entry, office systems and personal computing.

Networks include a broad range of cost/function trade-offs and technologies, in such diverse components as analog/digital converters, specialized and general-purpose terminals, line concentrators and multiplexers, communications links and low- to high-capacity processors.

The networking environment requires a master interconnection strategy so that these diverse products and applications can share computational and communications facilities while interacting compatibly.

Since its introduction in 1974, IBM's Systems Network Architecture has provided the blueprint by which the capabilities of IBM networking products have evolved in an orderly fashion. SNA provides rules for all levels of interaction, from physical/electrical interconnection of computing devices and terminals to meaningful application-oriented processing.

Thus one uniform design now eliminates the complexity and inefficiencies inherent when each type of product had to have its own specialized agreement with each other type. SNA is now integrated into the whole range of IBM products—from large mainframe computers to terminals to personal computers.

By eliminating the chaos once caused by incompatible implementations, SNA allows a computer user to communicate from office to office or from continent to continent.

An important feature of SNA is the organization of functions into multiple layers. In the most basic sense, different products can be configured into networks simply by adapting them to the transmission and electrical characteristics of the media interconnecting them. But physical interconnection does not result in meaningful communication. The lower layers control only the basic transfer of bits, while the higher layers support meaningful exchange of messages and documents and allow application-

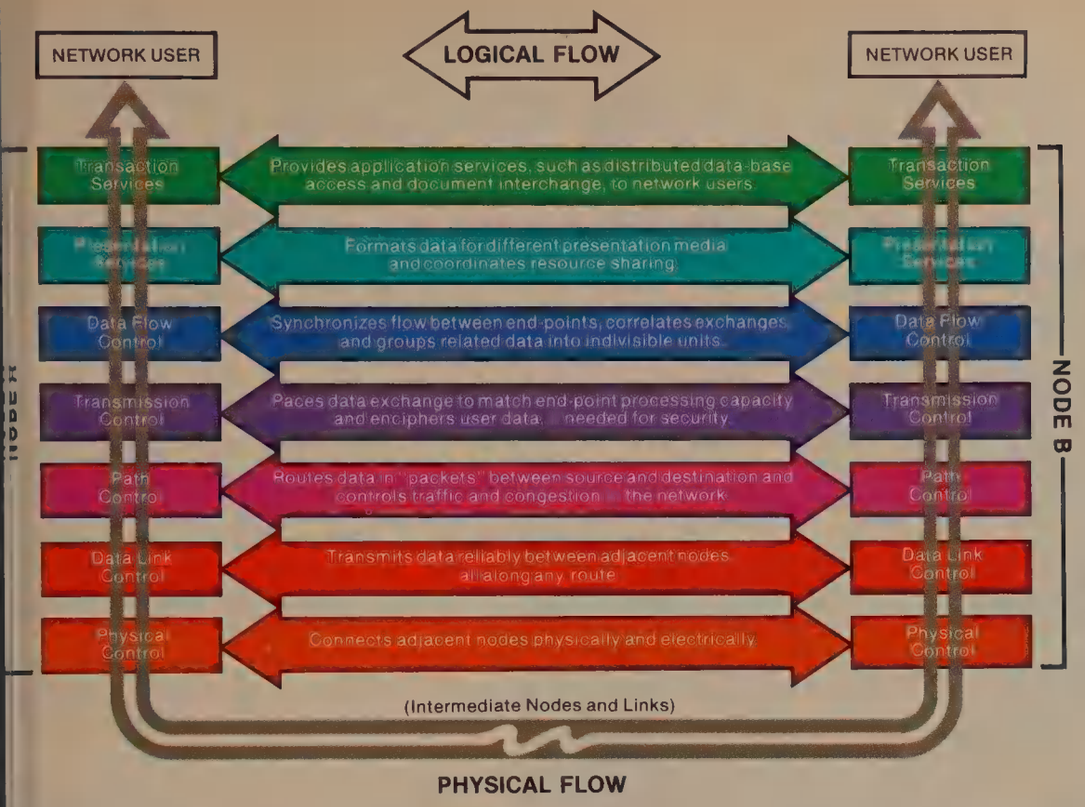


Figure 2. *Each node in an SNA network separates functions into multiple layers. Logically, a given layer in one node communicates with the corresponding layer in another node. This peer-to-peer communication relies on lower layers to transport the data.*

rogram interactions and data-base sharing. NA's separation of independent functions into nultiple layers means that changes in technolgy and capabilities can be confined to individal layers. This modular design eases daptation to network evolution.

SNA includes a variety of functions at ifferent layers of the architecture. For example, NA's Synchronous Data Link Control offers ncreased efficiency over earlier techniques. tate-of-the-art advances also have been made n traffic routing, congestion control and etwork availability. Additionally, SNA office ystems provide document encoding uniformity nd support distributed interchange, filing nd retrieval services.

SNA has also incorporated protocols dopted by national and international tandards organizations. This means SNA is ompatible with standards such as X.25 public acket switching, High-Level Data Link Control nd the Data Encryption Standard.

SNA management aids include product apabilities and software tools for planning, nstalling, changing, operating and maintaining networks. In today's environment, where annual growth and change typically can involve 20-50% of a network's facilities, aids such as these are critical to reduce operational expense and to foster optimal levels of network availability and performance.

IBM scientists, programmers and engineers around the world have spent collectively thousands of years of development on SNA. They continue to improve SNA's usability, manageability and performance, and also to extend its capabilities. Recent studies have focused on local-area networking, more dynamic reconfiguration within networks and interconnection of independent SNA networks.

SNA's success in reducing customer cost, while promoting ease of development of network applications, is reflected by a recent milestone —more than 10,000 large-system installations now incorporate SNA networking technology.

Systems Network Architecture is one example of IBM's commitment to product and technological leadership. Last year IBM's total worldwide investment in research, development and engineering was $3 billion.

IBM®

For free additional information on SNA, please write:
IBM Corporation, Dept. 605E/002
P.O. Box 12195, Research Triangle Park, NC 27709

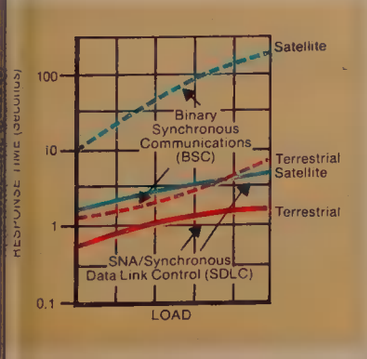


Figure 3. *Illustration of dramatic improvements in response time (using comparable display terminals) of SNA/SDLC over older data link controls such as BSC. For long-propagation-delay circuits, such as in satellite technology, the improvements in response time can be better than an order of magnitude.*



CENSORSHIP IN A FREE SOCIETY. IT'S A BAD MATCH.

Censorship is the greatest tragedy in American literature. It constricts the mind, teaches fear and leaves only ignorance and ashes.

Today, all over the country, books are being banned, burned and censored. Teachers, students, librarians, and book and magazine publishers are being harassed. The attacks of these self-appointed censors are endorsed by our silence.

The freedom to read is one of our most precious rights. Do something to protect it.

Contact:
People For The American Way, P.O. Box 19900
Washington, D.C. 20036 or call 202/822-9450.

People For The American Way

between the overlying sections of cord. It also depends on the diameter of the rod and cord, a geometry that affects the application of force by the top section to the bottom section. Bayman measured u and e for hitches of braided nylon string on a smooth steel rod; e was approximately 4, u about .2.

The clove hitch embodies both wraparounds and wrapovers. It is a common hitch whereby one can secure a cord to a fixture or to a heavier rope. From the free end the cord passes once around the rod and back over itself. Then it wraps once more around the rod before it emerges from under a second wrapover.

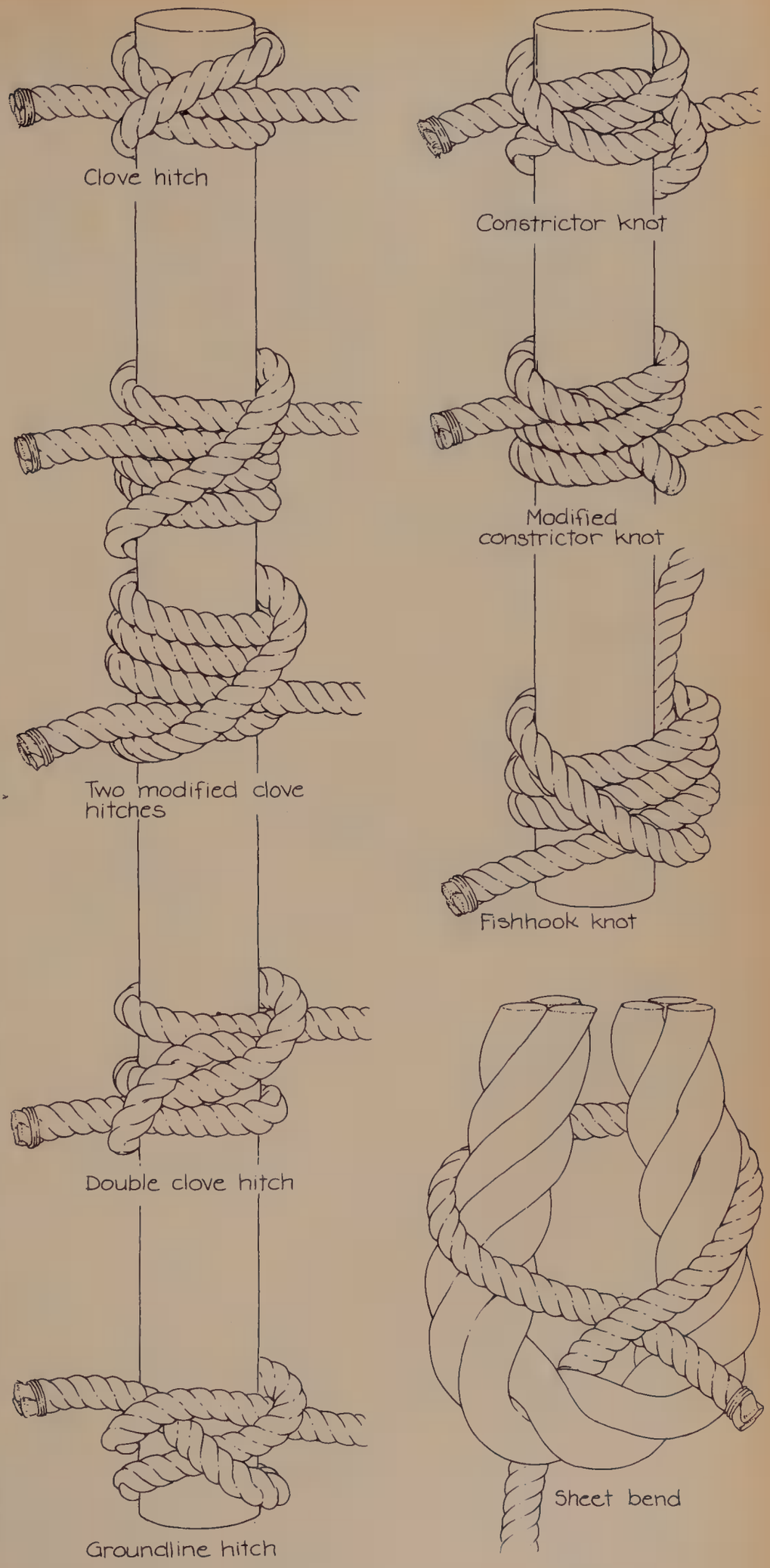
Sections of the cord are labeled in the bottom illustration at the left on the next page, beginning with 0 at the free end. Section 1 begins at the first wrapover and continues until the cord passes under the second wrapover. Section 2 emerges from the second wrapover and goes off to the load. (In application the hitch is pulled tighter than is shown.)

Is it a good hitch? Can the load on section 2 be arbitrarily large or will the load finally cause the free end to slip through the first wrapover and unravel the entire knot? Consider the bottom part of the first wrapover. On one side the free end has no tension. On the other side the cord has a tension (T_1) that acts to cause the free end to slide through the wrapover. The only opposing force is the friction generated by the top part of the wrapover where it presses the bottom part against the rod. If the hitch is to hold, T_1 cannot exceed the maximum possible value of this friction.

Next consider the two wraparounds of section 1. They do not slip if the tension in the part of the section near the second wrapover does not exceed $e^2 \times T_1$. Finally, consider the second wrapover. The cord will not slip under it if the tension (T_2) from the load does not overwhelm the opposing forces at the wrapover.

These relations of forces in the various parts of the hitch form a set of simultaneous inequalities that must be satisfied. The inequalities can be solved by a simple means of substitution, just as with simultaneous equations. The result is compact: all the conditions for avoiding slippage are satisfied if the value for u exceeds $1/e$. If this single condition is met, the load on the hitch can be arbitrarily large and the hitch will still hold. Additional load increases the pulls tending to loosen the knot but also increases the tensions that generate the opposing friction.

The first wrapover is the crucial element of the clove hitch. The force trying to pull the free end through the wrapover is equal to T_1. The opposing force is at most equal to u times the tension in the cord passing over the top of the wrapover. That tension is easily related to T_1 because only one wraparound lies



A variety of hitches

between the two points in the cord. Since the tension in the top part is at most eT_1, the friction at the wrapover is at most ueT_1. If u exceeds $1/e$, the hitch holds any arbitrarily large load on the cord. An increase in the load merely provides more friction to lock the hitch in place.

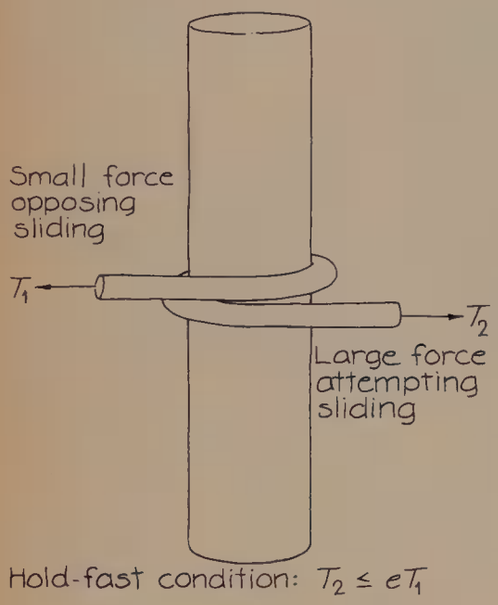
Would additional wraparounds in the clove hitch make it hold better in the sense that it can hold an arbitrarily large load with a smaller value of u? Two modified clove hitches are depicted in the top illustration on the opposite page. In the first hitch the cord is wrapped two more times around the rod before it crosses over the boundary between sections 0 and 1. In the second hitch the additional wraparounds are made just before the cord goes under its last wrapover and off to the load. These hitches are certainly more complex than the regular clove hitch, but are they stronger in the sense that they make less of a demand on the value of u? Is one of the modified hitches better than the other?

The key element in these modified clove hitches is the competition taking place near the free end. Consider the first knot. Section 1 is trying to make section 0 slide through the wrapover with tension T_1. The top section of the wrapover is opposing the slide by pressing on the cord.

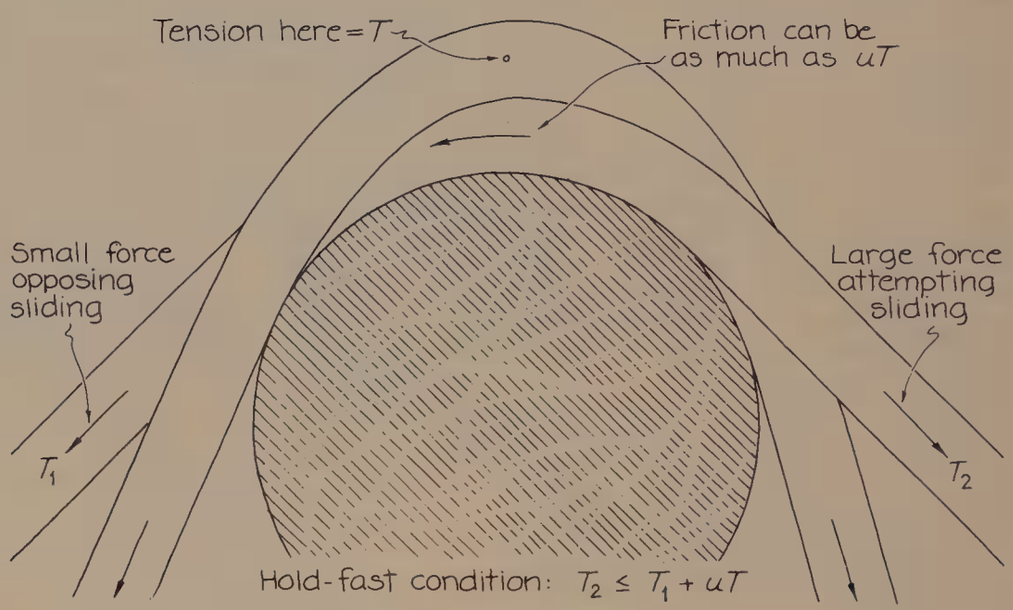
Does the cord slide? No, not if u is sufficiently large. Must it be as large as it is in the normal clove hitch? No, it can be considerably smaller because of the additional wraparounds. The tension in the top section of the first wrapover is now much larger than the force (T_1) tending to make the cord slide through the first wrapover. In the normal clove hitch the force from the wrapover can at most be ueT_1. The (implicit) power of 1 on the e is from the single wraparound between the top part of the wrapover and the place where T_1 pulls on the bottom part. In the modified hitch the force from the wrapover can be as much as ue^3T_1. Here the (explicit) power of 3 on the e comes from the three wraparounds between the top and bottom parts of the wrapover. Hence the modified hitch stands a better chance of holding fast.

How about the second modification? It appears to be as complex as the first, but again the key element lies in securing the free end. This time the competition of pulls is precisely like that in a normal clove hitch. The extra wraparounds do not serve to secure the free end and so give no improvement over the normal clove hitch. They do give more friction on the cord, thereby reducing the value of T_1 when the cord is put under load, but the hitch still must have u in excess of $1/e$ (just as it does in the clove hitch) if it is to hold.

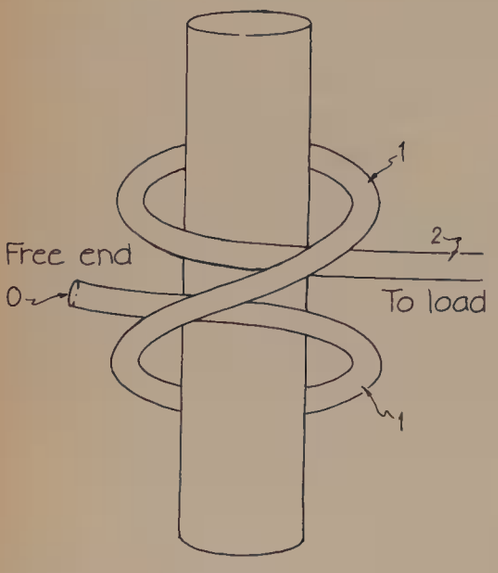
Something additional can be concluded from the first modified clove hitch. If it is to hold, the value of u must exceed $1/(e + e^2 + e^3)$. If u is larger than $1/(e + e^2)$, however, the first wraparound (section 1) serves no purpose and can be eliminated without jeopar-



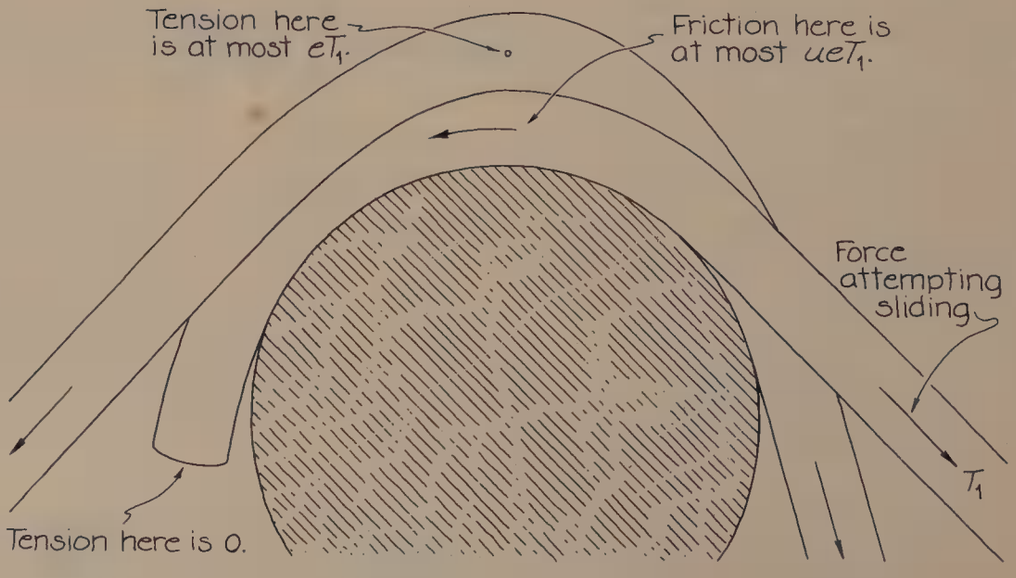
The forces in a wraparound



The forces in a wrapover



A clove hitch



The first wrapover of a clove hitch

dizing the integrity of the hitch. Section 0 might as well pass through a wrapover and become section 1. Furthermore, if u is larger than $1/e$, section 2 serves no purpose and can also be eliminated. Thus if u is larger than $1/e$, there is no reason to put extra wraparounds into a clove hitch because the hitch is already certain to hold fast.

Another possible modification is the double clove hitch depicted in the middle illustration at the left on this page. This hitch differs from the normal clove hitch because after the cord passes under the second wrapover it does not go directly to the load. Instead it makes another wrapover and another wraparound and then passes under itself to go to the load.

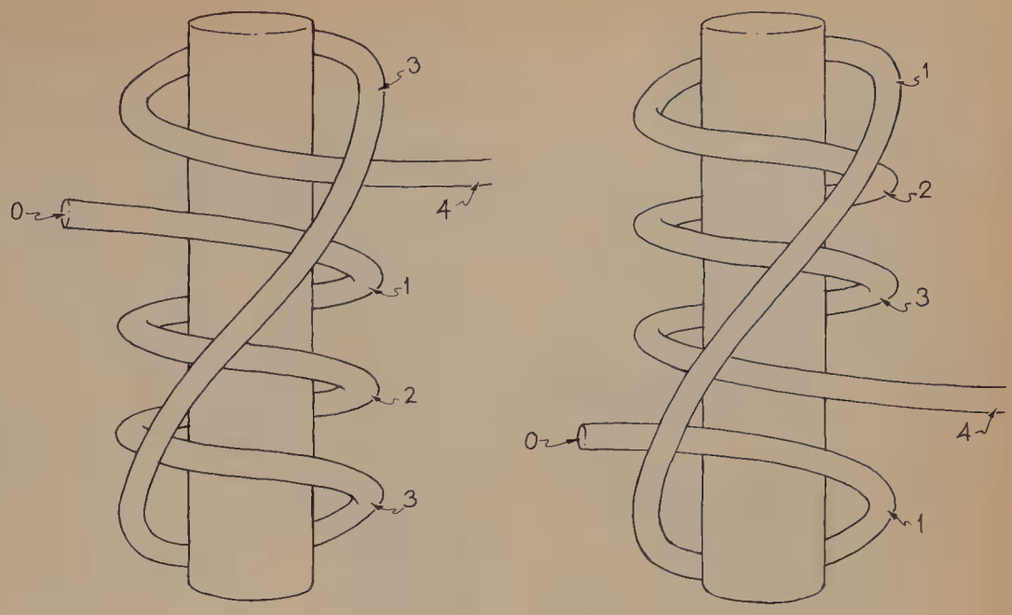
Does the extra hitching reduce the requirement on the value of u? No, at least not with the assumptions I have made about the primary sources of friction in knots. If the double clove hitch is to hold, u must still exceed $1/e$. The extra hitching serves primarily as a safety feature. If section 0 slips through its wrapover, you still have a clove hitch that must be untied before the cord loosens completely from the rod.

A knot similar to the clove hitch is the groundline hitch shown in the middle illustration at the right on this page. It differs from the clove hitch in the way the early stages of the knot are wrapped. Section 0 passes under a wrapover made by a late section in the knot, that is, a section near the one under full load. Section 1 wraps once around the rod and then passes under itself to begin section 2, which wraps once around the rod before it passes under the last wrapover and begins section 3.

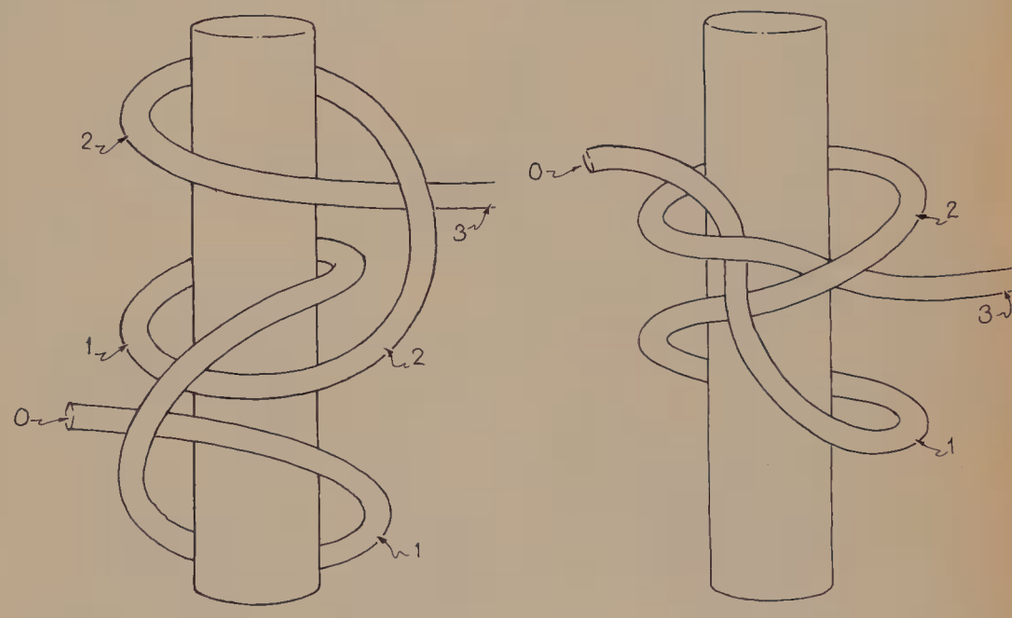
An analysis of the simultaneous inequalities indicates that this hitch holds fast if u and e meet the condition that $ue(u + e)$ exceed 1. Since e is greater than 1, the condition is met with a small value for u, smaller than the value of $1/e$ demanded by the clove hitch. Hence there is a range of values for u in which the groundline hitch holds but the clove hitch does not. If u is larger than $1/e$, both hitches serve equally well.

Why can the groundline hitch hold even when u is too small for the clove hitch to hold? The reason is found in the way the free end is tucked. In the clove hitch the free end is under the cord of section 1 after it has made one wraparound. In the groundline hitch the free end is tucked under a section of the cord closer to the region of full load. Therefore the tension in the top part of the wrapover that fixes the free end is larger than it is in the clove hitch. In general a hitch is stronger (in the sense that less is demanded of u) when the free end is tucked in this way.

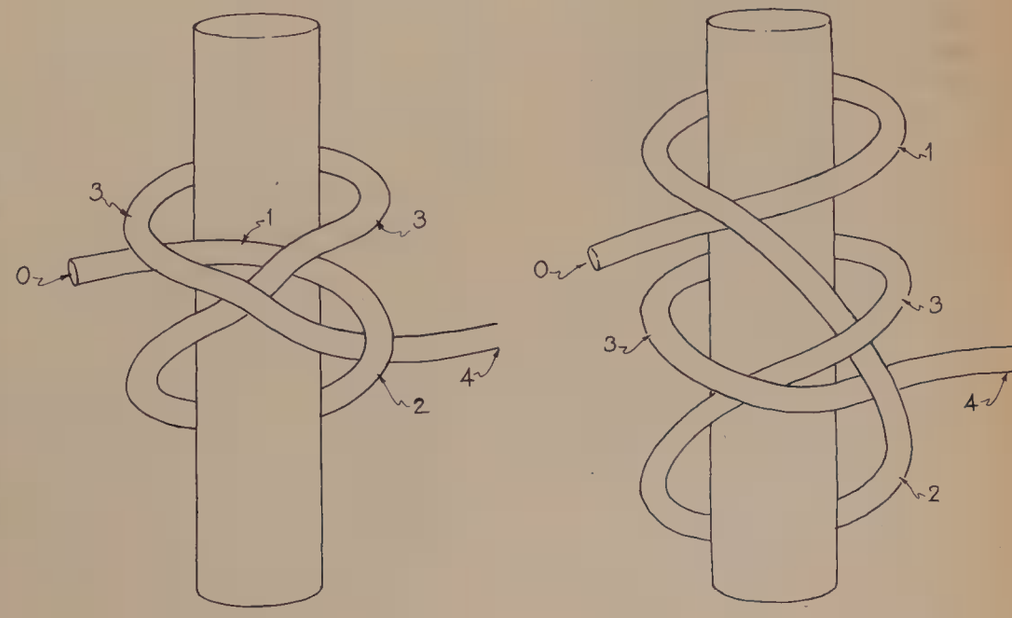
Bayman presents another example of how to tuck the free end beneficially. In the constrictor knot the free end of the



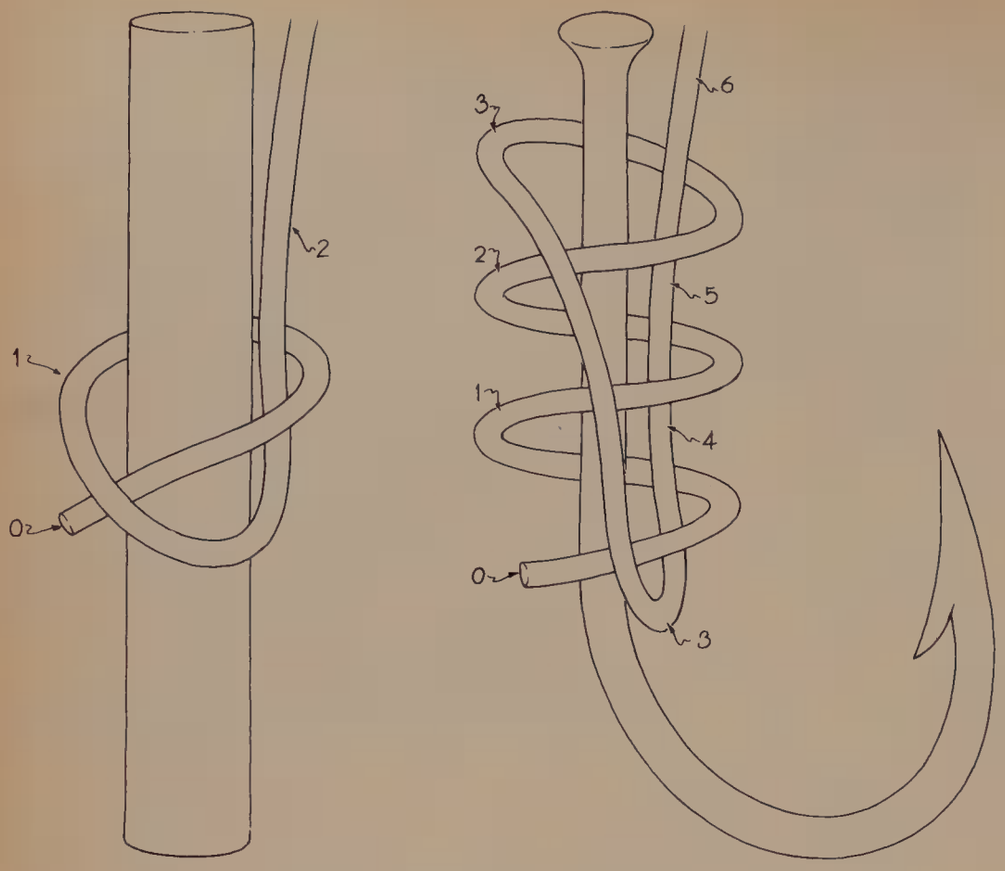
Modified clove hitches



A double clove hitch *A groundline hitch*



A constrictor knot *Benjamin F. Bayman's modified constrictor*



Knots for a fishhook

cord runs under two wrapovers, each of which is under large tension. The solution of the simultaneous inequalities yields the lower limit of u if the hitch is to hold fast: it must be larger than $1/(e(2 + e))$. Since this requirement is milder than the one for the clove hitch, the constrictor hitch is a better knot when you think the value for u may be low. If u is greater than $1/(2e)$, the friction in the wrapover marking the boundary between sections 1 and 2 is great enough so that the tension in section 1 can be zero and still the cord will not slip. If u is greater than $1/e$, section 2 is no longer required and can even have zero tension without jeopardizing the hitch. Hence the full constrictor hitch serves well if u is small compared with $1/e$, but the full hitch is not needed for larger values of u.

Bayman considers a modification of the constrictor hitch in which section 1 is wrapped around the rod once before it tucks under itself. Compare this knot with the full constrictor hitch. They differ only in the tying of section 1. In the regular hitch section 1 tucks under a section that is subject to a large tension, and so it is secure. In the modification section 1 gains some friction from the wraparound but then merely tucks under itself. Less tension presses down on the cord at the wrapover dividing sections 1 and 0.

This modified knot is certainly not as strong as the usual constrictor hitch. Is it better than a hitch in which section 1 is not tucked away at all? No, tucking section 1 under itself serves no purpose. In order for section 1 not to slip through the wrapover dividing it from section 2 the value of u must exceed $1/2e$.

Suppose it does not. Does the extra tie of section 1 save the hitch? No, because its wrapover on itself requires u to be larger than $1/e$ if the knot is not to slip. Therefore if the rest of the hitch begins to fail, the extra wraparound and wrapover of section 1 do not prevent the failure. You might as well leave them out, or better yet tie the full constrictor hitch.

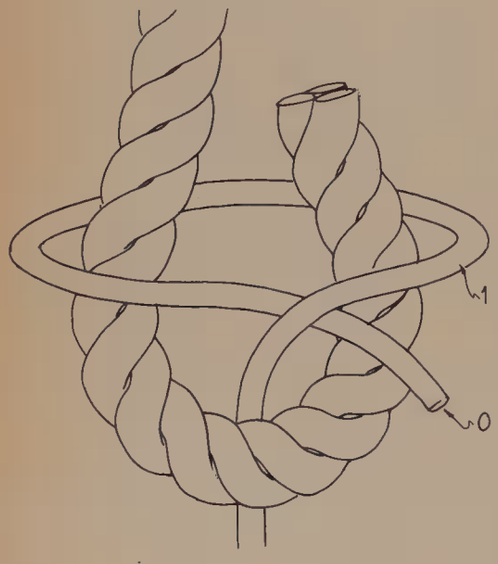
The top illustration on this page depicts a knot in which the cord leaves the knot parallel to the rod. This knot may be better than the clove hitch in some applications. I have not encountered a formal name for it; because it is similar to a knot for securing a fishing line to a straight hook (one without an eye the line can be tied to), I call it a fishhook knot.

When I solve the simultaneous inequalities for this knot, I find that (within the assumptions I have been applying) the hook knot demands the same kind of value for u as the clove hitch. In both hitches section 1 tries to pull the free end through a wrapover. Also in both knots the tension in the top section of the wrapover is T_1 multiplied by the factor due to one wraparound and thus is at most eT_1. If the cord is to remain in place, the value of u must exceed $1/e$. Although the two knots look different and may serve different functions in application, they are identical in their ability to withstand an arbitrarily large load.

If a cord or a fishing line is to be tied to a straight hook, the cord ought to be wrapped around the shaft of the hook several times before it is wormed through several wrapovers. How effective are the extra wraparounds? Does adding one more wraparound make a significant change in the demand on the value for u?

When I solved the simultaneous inequalities for this knot, I found that the minimum value of u is driven downward sharply with each additional wraparound. For the knot shown the value for u must exceed $1/(e + e^2 + e^3)$. If you eliminate one of the wraprounds, the term e^3 drops out of the formula. Since e is larger than 1, the loss of one wraparound can have a significant effect.

The bottom illustration at the left on this page shows a sheet-bend hitch designed for joining a cord of small diameter to a larger rope. The rope forms a *U*. From the free end the cord wraps



A sheet-bend hitch

	$j=1$	$j=2$	$j=3$
$i=1$	1	0	$-ue$
$i=2$	-1	1	$-u$
$i=3$	0	$-e$	$1-ue$

$$\text{Determinant} = 1-ue(2+e)=0$$

$$u=\frac{1}{e(2+e)} \text{ for critical case}$$

Constrictor-knot determinant

around the full U once and then back over itself before it passes through the bottom of the U and off to the load.

The "lock" on this knot is in the wrapover. In the top part the cord has a large tension, almost as large as in the section under full load. The bottom part of the wrapover is therefore pressed hard against the rope. If the wrapover fails, the entire sheet bend fails.

To analyze the sheet bend's friction requirement I make several assumptions. First I assume the rope is fairly stiff so that in the wrapover the bottom part can be squeezed between the rope and the top part. I also assume the wraparound on the U is identical with one around a rod of large diameter. Within the limits of these assumptions the requirement on u for the cord to hold is that it must exceed $1/e$, just as with a clove hitch. The sheet bend is in the same class as a clove hitch.

The sheet bend can be improved by wrapping the cord around the U more times before making a wrapover and passing the cord through the bottom of the U. Each added wraparound increases the effectiveness of the wrapover by widening the difference in tension between the top and bottom parts of the wrapover.

All the knots I have discussed can be investigated by solving simultaneous inequalities. Bayman has devised a quicker solution in which a determinant is constructed from the coefficients in the inequality statements. If the quantities of e and u lie at the crucial values where the cord is on the verge of slipping, the simultaneous statements are equations and not inequalities. Hence at the crucial values the determinant is equal to zero. Establish it at zero and then expand it in the usual way. The resulting equation is solved for u. This value of u must be exceeded if the knot is to hold under load.

The first step in constructing a determinant is to section the hitch for analysis. The free end is labeled 0 as before. Each of the succeeding sections is defined as a length of cord that begins at the bottom of a wrapover and extends until it must pass under another one. The determinant is laid out on a chart labeled with i vertically and j horizontally. These letters represent the various sections in the hitch. Neither section 0 nor the last section (the one under the full load) is included.

The determinant is filled in by horizontal rows. Within a row the spaces under the values of j are determined through three rules: (1) If j matches i, the space has a term of 1; (2) if j is the section where i begins, the space has a term of $-ue^n$, where n is the number of wraparounds j makes before it passes over the beginning of i; (3) if j is the section just before i in the hitch, the space has a term of $-ue^n$, where n represents the total number of wraparounds the previous section has.

The determinant for the full constrictor hitch is worked out in the bottom illustration at the right on the opposite page. The top row is for section 1, the middle row for section 2 and the bottom row for section 3. In the top row j and i match in the first space, and so a 1 is inserted. The other rules do not contribute any further terms to the space. They also do not contribute anything to the space below $j = 2$, which is left at zero.

The space below $j = 3$ does have a contribution because section 1 begins under section 3. Since section 3 has been through one full wraparound when it crosses over the beginning of section 1, the value of n in the term contributed by the second rule is 1.

The second row pertains to section 2 of the hitch. The space below $j = 2$ has a value of 1 because of the match of i and j. The space below $j = 1$ deals with the section of the hitch just ahead of section 2. Section 1 had no full wraparounds, and so the term from rule 3 that is to be filled in is $-e^0$, or -1.

The space below $j = 3$ relates to the beginning of section 2. Since it begins under section 3 before that section has made a wraparound, the contribution to the determinant is $-ue^0$, a term that reduces to $-u$.

The bottom row in the determinant pertains to section 3 of the hitch. The space below $j = 1$ has no contribution. The space below $j = 2$ has $-e$ because section 3 is preceded by section 2, which has one wraparound. The space below $j = 3$ has two contributions based on the three rules. The match of j and i contributes a term of 1. A second term is contributed because section 3 begins under itself after it has made one wraparound. This second term is $-ue^1$, which simplifies to $-ue$.

Once the determinant is constructed it is set equal to zero and solved by multiplying terms along diagonals or by reducing the determinant to smaller determinants. The resulting equation involves both u and e. The solution for u in the equation is the critical value for u. If u is smaller than this value, the cord slips because the wrapovers will not hold under tension. If u is larger than the critical value, the cord can be expected to hold under load.

A great deal more can be done on the analysis of hitches and other knots. Besides investigating new knots or improving old ones you might like to strengthen the theoretical analysis. For example, Bayman points out that not all wrapovers are identical in their ability to squeeze an underlying section of cord, particularly when one section of cord forms the top part of several adjacent wrapovers. If you do strengthen the theory or if you find interesting new knots, I would enjoy hearing from you.



NordicTrack

Jarless Total Body Cardiovascular Exerciser

Duplicates X-C Skiing For The Best Way To Fitness

NordicTrack duplicates the smooth rhythmic total body motion of XC Skiing. Recognized by health authorities as the most effective fitness building exercise available. Uniformly exercises more muscles than jogging, swimming, cycling and rowing.

Does Not cause joint or back problems as in jogging. Highly effective for weight control and muscle toning.

Easily Adjustable for arm resistance, leg resistance and body height. Smooth, quiet action. Folds compactly to require only 15 by 17 inches of storage area. Lifetime quality.

Used In thousands of homes and many major health clubs, universities, and corporate fitness centers.

Call or Write for FREE BROCHURE
Toll Free 1-800-328-5888 MN 612-448-6987
PSI 124 F Columbia Crt., Chaska, Minn. 55318

SCIENTIFIC AMERICAN

is now available to the blind and physically handicapped on cassette tapes.

All inquiries should be made directly to RECORDED PERIODICALS, Division of Volunteer Services for the Blind, 919 Walnut Street, 8th Floor, Philadelphia, PA 19107.

ONLY the blind or handicapped should apply for this service. There is a nominal charge.

BIBLIOGRAPHY

Readers interested in further explanation of the subjects covered by the articles in this issue may find the following lists of publications helpful.

MATHEMATICAL GAMES

ON WELL-QUASI-ORDERING FINITE TREES. C. St. J. A. Nash-Williams in *Proceedings of the Cambridge Philosophical Society,* Vol. 59, Part 4, pages 833–835; October, 1963.

NUMBER THEORY: THE THEORY OF PARTITIONS. George E. Andrews. Addison-Wesley Publishing Co., 1976.

TREES AND BALL GAMES. Raymond M. Smullyan in *Annals of the New York Academy of Sciences,* Vol. 321, pages 86–90; 1979.

PROVING TERMINATION WITH MULTISET ORDERINGS. Nachum Dershowitz and Zohar Manna in *Communications of the ACM,* Vol. 22, No. 8, pages 465–476; August, 1979.

ACCESSIBLE INDEPENDENCE RESULTS FOR PEANO ARITHMETIC. Laurie Kirby and Jeff Paris in *The Bulletin of the London Mathematical Society,* Vol. 14, No. 49, Part 4, pages 285–293; July, 1983.

TRAUMA

THE HEALING HAND: MAN AND WOUND IN THE ANCIENT WORLD. Guido Majno. Harvard University Press, 1975.

THE EPIDEMIOLOGY AND PREVENTION OF INJURIES. Susan P. Baker and P. E. Dietz in *The Management of Trauma,* edited by George D. Zuidema, Robert B. Rutherford and Walter F. Ballinger II. W. B. Saunders Company, 1979.

INJURY CONTROL. W. Haddon and S. P. Baker in *Preventive and Community Medicine.* Duncan W. Clark and Brian MacMahon. Little, Brown & Co., 1981.

TRAUMA MANAGEMENT, VOL. 1: ABDOMINAL TRAUMA. Edited by F. William Blaisdell and Donald D. Trunkey. Thieme-Stratton, Inc., New York, 1982.

THE PURIFICATION AND MANUFACTURE OF HUMAN INTERFERONS

VIRUS INTERFERENCE, 1: THE INTERFERON. A. Isaacs and J. Lindenmann in *Proceedings of the Royal Society of London, Series B,* Vol. 147, No. 927, pages 258–267; September 12, 1957.

THE INTERFERON SYSTEM. Edited by W. E. Stewart. Springer-Verlag, 1979.

INTERFERONS. Edited by S. Pestka in *Methods in Enzymology,* Part A, Vol. 78; Part B, Vol. 79; 1981.

AMINO ACID SEQUENCE OF A HUMAN LEUKOCYTE INTERFERON. Warren P. Levy, Menachem Rubinstein, John Shively, Ursino Del Valle, Chen-Yen Lai, John Moschera, Larry Brink, Louise Gerber, Stanley Stein and Sidney Pestka in *Proceedings of the National Academy of Sciences of the United States of America,* Vol. 78, No. 10, pages 6186–6190; October, 1981.

THE INTERFERON SYSTEM: A REVIEW TO 1982. Samuel Baron, Ferdinando Dianzani and G. John Stanton in *Texas Reports on Biology and Medicine,* Vol. 41, Part 1, Part 2; 1982.

THE HUMAN INTERFERONS—FROM PROTEIN PURIFICATION AND SEQUENCE TO CLONING AND EXPRESSION IN BACTERIA: BEFORE, BETWEEN, AND BEYOND. Sidney Pestka in *Archives of Biochemistry and Biophysics,* Vol. 221, No. 1, pages 1–37; February 15, 1983.

MAGNETIC FIELDS IN THE COSMOS

THE SOLAR WIND. E. N. Parker in *Scientific American,* Vol. 210, No. 4, pages 66–76; April, 1964.

COSMICAL MAGNETIC FIELDS: THEIR ORIGIN AND THEIR ACTIVITY. E. N. Parker. Oxford University Press, 1979.

MAGNETIC STARS. E. F. Borra, J. D. Landstreet and L. Mestel in *Annual Review of Astronomy and Astrophysics,* Vol. 20, pages 191–220; 1982.

INTERSTELLAR MATTER IN METEORITES

ISOTOPIC ANOMALIES IN METEORITES. F. Begemann in *Reports on Progress in Physics,* Vol. 43, Part 4, pages 1309–1356; 1980.

COMPOUNDS IN METEORITES AND THEIR ORIGINS. R. Hayatsu and E. Anders in *Topics in Current Chemistry,* Vol. 99, pages 1–34; 1981.

NOBLE GASES IN METEORITES: EVIDENCE FOR PRESOLAR MATTER AND SUPERHEAVY ELEMENTS. E. Anders in *Proceedings of the Royal Society of London, Series A,* Vol. 374, No. 1757, pages 207–238; February 4, 1981.

INTERSTELLAR CARBON IN METEORITES. P. Swart, M. Grady, C. Pillinger, R. Lewis and E. Anders in *Science,* in press.

THE CHEMICAL DEFENSES OF TERMITES

CHEMICAL DEFENSE BY TERMITE SOLDIERS. Glenn D. Prestwich in *Journal of Chemical Ecology,* Vol. 5, pages 459–480; 1979.

CHEMISTRY, DEFENSE AND SURVIVAL: CASE STUDIES AND SELECTED TOPICS. Thomas Eisner in *Insect Biology in the Future,* edited by Michael Locke and David S. Smith. Academic Press, 1980.

CHEMICAL SELF-DEFENSE BY TERMITE WORKERS: PREVENTION OF AUTOTOXICATION IN TWO RHINOTERMINIDS. Stephen G. Spanton and Glenn D. Prestwich in *Science,* in press.

RATIONAL COLLECTIVE CHOICE

SOCIAL CHOICE AND INDIVIDUAL VALUES. Kenneth J. Arrow. John Wiley & Sons, Inc., 1963.

AGGREGATION OF PREFERENCES. Donald J. Brown in *Quarterly Journal of Economics,* Vol. 89, No. 3, pages 456–469; August, 1975.

CHOICE, WELFARE AND MEASUREMENT. Amartya Sen. The MIT Press, 1982.

ACYCLIC COLLECTIVE CHOICE RULES. Douglas H. Blair and Robert A. Pollak in *Econometrica,* Vol. 50, No. 4, pages 931–943; July, 1982.

THE STAVE CHURCHES OF NORWAY

STAV OG LAFT I NORGE/EARLY WOODEN ARCHITECTURE IN NORWAY. Gunnar Bugge and Christian Norborg-Schulz. Byggekunst, Norske Arkitekters Landsforbund, Oslo, 1969.

NORWEGIAN STAVE CHURCHES. Roar Hauglid. English text by R. I. Christophersen. Dreyer, Oslo, 1970.

THE STAVE CHURCHES OF KAUPANGER. Kristian Bjerkens and Hans-Emil Lidén. Fabritius Forlag, Oslo, 1975.

NORSKE STAVKIRKER: BYGNINGSHISTORISK BAKGRUNN OG UTVIKLING. Roar Hauglid. Dreyer, Oslo, 1976.

DIGITAL TYPOGRAPHY

AN ATLAS OF TYPEFORMS. James Sutton and Alan Bartram. Hastings House, 1968.

TRANSMISSION AND DISPLAY OF PICTORIAL INFORMATION. D. E. Pearson. Pentech Press, Ltd., 1975.

ANCIENT WRITING AND ITS INFLUENCE. Berthold Louis Ullman. University of Toronto Press, 1980.

CHARACTER GENERATION BY COMPUTER. Ph. Coueignoux in *Computer Graphics and Image Processing,* Vol. 16, pages 240–269; 1981.

THE CONCEPT OF A META-FONT. Donald E. Knuth in *Visible Language,* Vol. 16, No. 1, pages 3–27; Winter, 1982.

DIGITAL TYPEFACE DESIGNS. Charles Bigelow. Seybold Publications, 1983.

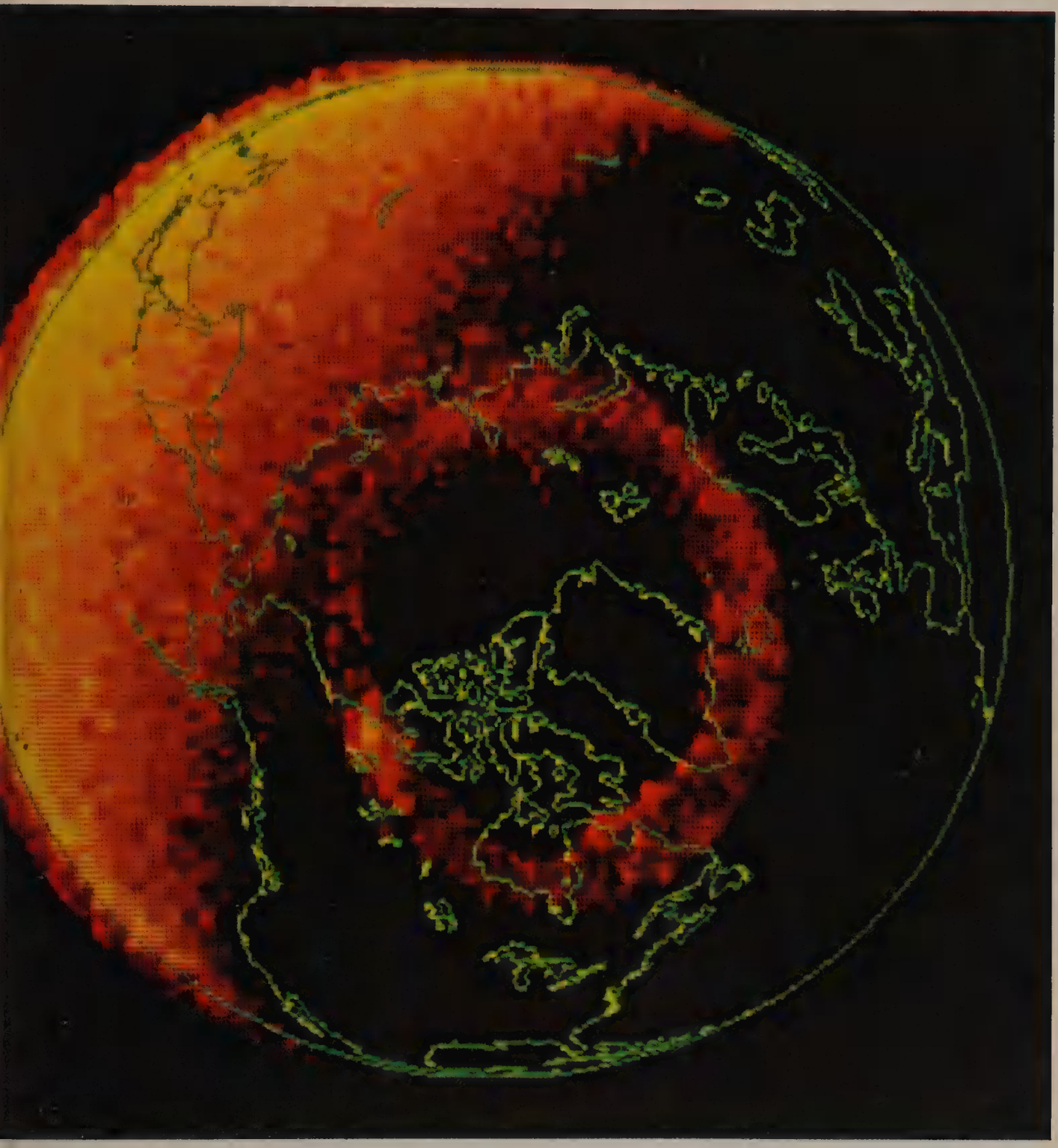
THE AMATEUR SCIENTIST

THEORY OF HITCHES. Benjamin F. Bayman in *American Journal of Physics,* Vol. 45, No. 2, pages 185–190; February, 1977.

SEP 13 '88

Marygrove College Library
Detroit, Michigan 48221
PLEASE DO NOT REMOVE

SCIENTIFIC AMERICAN



THE DYNAMIC EARTH

$2.50

September 1983

0 74820 08715 09

How Exxon has pioneered systems to optimize

Roy Lieber works at the leading edge of process control.



As an operator at an Exxon refinery monitors the schematic of a process on the CRT, the representation of the furnace suddenly changes color and begins to pulse, indicating a condition requiring attention. He touches the screen and corrective action is taken. Elsewhere another operator advises the system by CRT to switch a tower's emphasis from propane to butane as a result of analysis of market data.

Such advanced process control is the result of a long history of pioneering in the application of computers to refinery operation. It goes back to the early 1960's when Roy Lieber and his colleagues at Exxon Research and Engineering Company (ER&E) introduced the use of computers for closed loop control. These first systems, which used the primitive minicomputers of that era, provided superior regulatory control despite their limited functionality.

Precision refining

As computer technology advanced, the systems designed by ER&E became progressively more sophisticated. Combining control theory with process know-how further expanded applications, including the automation of related refinery functions such as blending, product storage and shipping.

Current systems permit implementation of the complex control strategies that maximize product yields and minimize energy consumption while meeting other product and environmental objectives.



advanced computer refinery operation.

User friendliness is also being stressed. Through an interconnected, hierarchical network of micro, mini and maxi computers, today's operators monitor the entire refinery from work stations housing three to five CRTs.

Future trends

Exxon has installed more than 100 closed loop computer control systems in its refineries and chemical plants worldwide. In fact, the majority of these plants are totally under computer control.

Advanced as the computerized process control systems are today, Roy Lieber feels there is still great potential for future economies and efficiencies. He talks about work on integrated networks of systems sharing data on all aspects of refinery operation both within and between plants. Using advances in electronics and concepts such as artificial intelligence, the goal is to optimize single refineries and conceivably even refineries on a regional and a worldwide basis. He is helping guide the development of the hardware and software that will make this possible.

Exxon Research and Engineering Company

Process control technology is but one example of the numerous activities underway at Exxon Research and Engineering Company. A wholly owned subsidiary of Exxon Corporation, ER&E employs more than 2,000 scientists and engineers working on petroleum products and processing, synthetic fuels, pioneering science and the engineering required to develop and apply new technologies in the manufacture of fuels and other products. For more information on process control or ER&E, write Dr. E.E. David, Jr., President, Exxon Research and Engineering Company, Room 705, P.O. Box 101, Florham Park, N.J. 07932.

EXXON



The majestic Wild Turkey–a fitting symbol of our country's finest native whiskey–can be seen in the woods bordering our distillery.

Wild Turkey Hill. A place unlike any other.





The woods on Wild Turkey Hill slope down to the edge of the Kentucky River. On top of the hill, there's been a distillery for nearly 150 years. It's a unique spot: gently running waters below and constant breezes above that cool our Wild Turkey whiskey naturally as it ages in the barrel. Wild Turkey Hill is a very special place. And it helps us make Wild Turkey very special.

WILD TURKEY®/101 PROOF/8 YEARS OLD
AUSTIN, NICHOLS DISTILLING CO., LAWRENCEBURG, KENTUCKY© 1983

Established 1845 **SCIENTIFIC AMERICAN** September 1983 Volume 249 Number 3

ARTICLES

46 **THE DYNAMIC EARTH, by Raymond Siever**
Introducing an issue about the earth as a dynamic system that maintains itself in a steady state.

56 **THE EARTH'S CORE, by Raymond Jeanloz**
In it the turbulent flow of liquid iron alloyed with other elements generates the geomagnetic field.

66 **THE EARTH'S MANTLE, by D. P. McKenzie**
Convection currents in its ductile silicates drive the motions of the overlying lithospheric plates.

114 **THE OCEANIC CRUST, by Jean Francheteau**
It is created and destroyed in a flow outward from the midocean ridges and back into the mantle.

130 **THE CONTINENTAL CRUST, by B. Clark Burchfiel**
Much older than oceanic crust, it is steadily reworked by tectonics, volcanism, air and water.

146 **THE OCEAN, by Wallace S. Broecker**
Its dynamic behavior is exemplified by long-term shifts in the substances that are dissolved in it.

162 **THE ATMOSPHERE, by Andrew P. Ingersoll**
This low-viscosity fluid acts to rapidly distribute solar energy that is intercepted by the earth.

176 **THE BIOSPHERE, by Preston Cloud**
Life powerfully shapes the evolution of the lithosphere, the hydrosphere and the atmosphere.

DEPARTMENTS

6 LETTERS
10 50 AND 100 YEARS AGO
15 THE AUTHORS
18 MATHEMATICAL GAMES
37 BOOKS
82 SCIENCE AND THE CITIZEN
190 THE AMATEUR SCIENTIST
202 BIBLIOGRAPHY

BOARD OF EDITORS Gerard Piel (Publisher), Dennis Flanagan (Editor), Brian P. Hayes (Associate Editor), Philip Morrison (Book Editor), John M. Benditt, Peter G. Brown, Michael Feirtag, Diana Lutz, Jonathan B. Piel, John Purcell, James T. Rogers, Armand Schwab, Jr., Joseph Wisnovsky

ART DEPARTMENT Samuel L. Howard (Art Director), Steven R. Black (Assistant Art Director), Ilil Arbel, Edward Bell

PRODUCTION DEPARTMENT Richard Sasso (Production Manager), Carol Eisler and Leo J. Petruzzi (Assistants to the Production Manager), Carol Hansen (Electronic Composition Manager), Carol Albert, Karen Friedman, Lori Mogol, Karen O'Connor, Julio E. Xavier

COPY DEPARTMENT Sally Porter Jenks (Copy Chief), Debra Q. Bennett, Mary Knight, Dorothy R. Patterson

GENERAL MANAGER George S. Conn

ADVERTISING DIRECTOR C. John Kirby

CIRCULATION MANAGER William H. Yokel

SECRETARY Arlene Wright

SCIENTIFIC AMERICAN (ISSN 0036-8733), PUBLISHED MONTHLY BY SCIENTIFIC AMERICAN, INC., 415 MADISON AVENUE, NEW YORK, N.Y. 10017. COPYRIGHT © 1983 BY SCIENTIFIC AMERICAN, INC. ALL RIGHTS RESERVED. PRINTED IN THE U.S.A. NO PART OF THIS ISSUE MAY BE REPRODUCED BY ANY MECHANICAL, PHOTOGRAPHIC OR ELECTRONIC PROCESS, OR IN THE FORM OF A PHONOGRAPHIC RECORDING, NOR MAY IT BE STORED IN A RETRIEVAL SYSTEM, TRANSMITTED OR OTHERWISE COPIED FOR PUBLIC OR PRIVATE USE WITHOUT WRITTEN PERMISSION OF THE PUBLISHER. SECOND-CLASS POSTAGE PAID AT NEW YORK, N.Y., AND AT ADDITIONAL MAILING OFFICES. AUTHORIZED AS SECOND-CLASS MAIL BY THE POST OFFICE DEPARTMENT, OTTAWA, CANADA, AND FOR PAYMENT OF POSTAGE IN CASH. SUBSCRIPTION RATE: $24 PER YEAR IN THE U.S. AND ITS POSSESSIONS; $33 PER YEAR IN ALL OTHER COUNTRIES. POSTMASTER: SEND ADDRESS CHANGES TO SCIENTIFIC AMERICAN, 415 MADISON AVENUE, NEW YORK, N.Y. 10017.

SCIENTIFIC AMERICAN

CORRESPONDENCE

Offprints of more than 1,000 selected articles from earlier issues of this magazine, listed in an annual catalogue, are available at 75 cents each. Correspondence, orders and requests for the catalogue should be addressed to W. H. Freeman and Company, 4419 West 1980 South, Salt Lake City, UT 84121. Offprints adopted for classroom use may be ordered direct or through a college bookstore. Sets of 10 or more Offprints are collated by the publisher and are delivered as sets to bookstores.

Photocopying rights are hereby granted by Scientific American, Inc., to libraries and others registered with the Copyright Clearance Center (CCC) to photocopy articles in this issue of SCIENTIFIC AMERICAN for the flat fee of $2 per copy of each article or any part thereof. Such clearance does not extend to the photocopying of articles for promotion or other commercial purposes. Correspondence and payment should be addressed to Copyright Clearance Center, Inc., 21 Congress Street, Salem, MA 01970. Specify CCC Reference Number ISSN 0036–8733/83. $2.00 + 0.00.

Editorial correspondence should be addressed to The Editors, SCIENTIFIC AMERICAN, 415 Madison Avenue, New York, NY 10017. Manuscripts are submitted at the authors' risk and will not be returned unless they are accompanied by postage.

Advertising correspondence should be addressed to C. John Kirby, Advertising Director, SCIENTIFIC AMERICAN, 415 Madison Avenue, New York, NY 10017.

Subscription correspondence should be addressed to Subscription Manager, SCIENTIFIC AMERICAN, P.O. Box 5969, New York, NY 10017. The date of the last issue on your subscription is shown in the upper right-hand corner of each month's mailing label. For change of address notify us at least four weeks in advance. Please send your old address (if convenient, on a mailing label of a recent issue) as well as the new one.

Name

New Address

Street

City

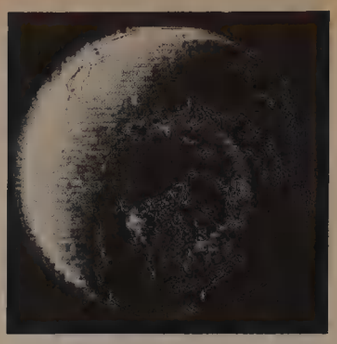
State and ZIP

Old Address

Street

City

State and ZIP



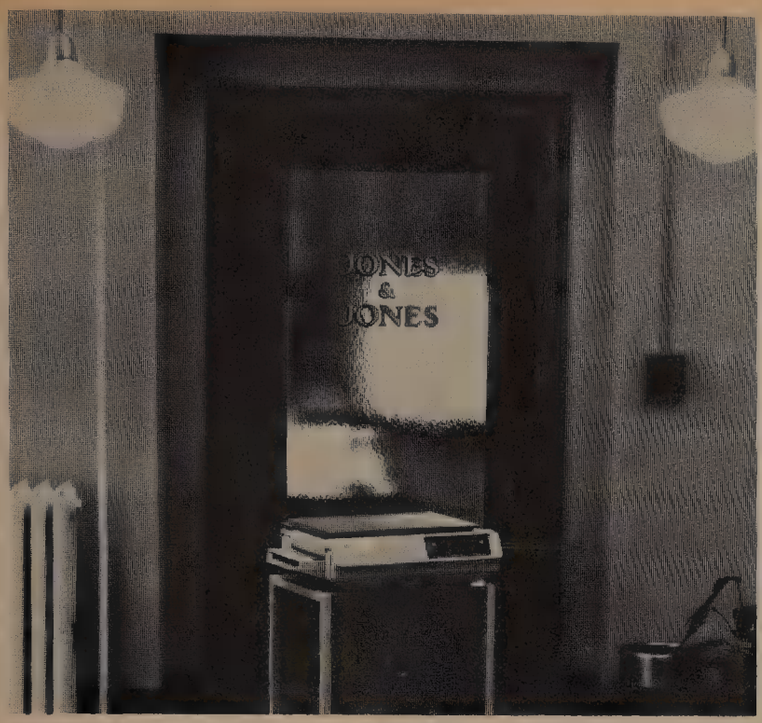
THE COVER

The computer-graphics image on the cover symbolizes the theme of this issue of SCIENTIFIC AMERICAN: the dynamic earth. The image shows an aurora over the earth's north magnetic pole viewed from a satellite at an altitude of 3.19 earth radii. The aurora is the luminous ellipse projected on the green computer-generated map of the globe. The bright crescent at the upper left is the illuminated daylight side of the earth. Auroras are emitted by molecules in the earth's upper atmosphere excited by electrically charged particles in the "wind" of thin gas expanding outward from the sun. The particles plunge into the atmosphere because they are trapped in the earth's magnetic field, which is generated by slow flows in the earth's metallic core. Therefore the aurora is a manifestation of the dynamic activity of the core. The image was made by a University of Iowa team from data transmitted by instrumentation aboard the satellite *Dynamics Explorer 1*. It is provided through the courtesy of Louis A. Frank.

THE ILLUSTRATIONS

Cover courtesy of Louis A. Frank, University of Iowa

Page	Source
18–28	Albert E. Miller
47	By permission of the Houghton Library, Harvard University
48–49	Sally Black
51	Christopher S. Scotese
52–55	Sally Black
57	Louis A. Frank
58–65	George V. Kelvin
66	Flagstaff Image Processing Center, U.S. Geological Survey
68–73	Ian Worpole
74	F. R. Boyd (*top left and third from left*), Euan G. Nisbet (*second from left*), Lawrence Grossman (*top right*), Ian Worpole (*bottom*)
75–77	Ian Worpole
78	EROS Data Center (*left*), Ilil Arbel (*right*)
115–117	William F. Haxby
118–119	Andrew Tomko
120	Todd Pink
121	William B. F. Ryan
122	Jean Francheteau
126–127	Todd Pink
128	Peter Buhl
129	Todd Pink
131	EROS Data Center
132–133	Andrew Tomko
134–138	Todd Pink
139	B. Clark Burchfiel
140–142	Todd Pink
147	National Aeronautics and Space Administration
148–154	Alan D. Iselin
156–157	Andrew Tomko
158–160	Alan D. Iselin
163	Laboratory for Atmospheric and Space Physics, University of Colorado at Boulder
164	Walken Graphics
165–167	Jet Propulsion Laboratory, California Institute of Technology
168–169	European Centre for Medium Range Weather Forecasts, Reading, England
170–174	Walken Graphics
177	EROS Data Center
178	Preston Cloud
179–181	Patricia J. Wynne
183	Courtesy of Norman D. Newell, American Museum of Natural History
185	Preston Cloud
188	Patricia J. Wynne
191	Jearl Walker
195–201	Michael Goodman









AT 3M, WE FIND IT RELATIVELY SIMPLE TO KEEP UP WITH THE JONESES.

In this fast-paced, workaday world, keeping up with any office's copying needs is no small task.

But 3M's line of copiers is big enough to answer the needs of virtually every business. Large or small.

From our compact Parabond models to our versatile plain paper models, 3M copiers offer a wide array of features and options, including microprocessing, collating, reduction, edge-to-edge copying and liquid crystal display.

For the full story on 3M copiers, mail us this coupon. Or call toll-free 800-328-1684. (In Minnesota: 800-792-1072).

Because if we can keep up with the Joneses, we can certainly keep up with you.

3M hears you...

Please send me more information on the 3M line of copiers. SCA 9/83

Name _____

Title _____

Company _____

Phone _____

Address _____ **City** _____

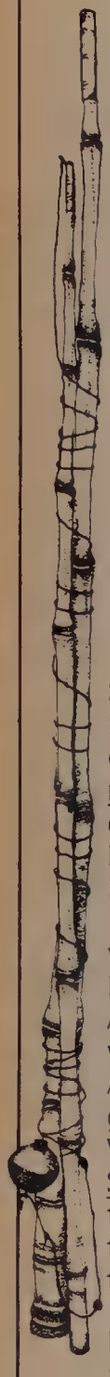
State _____ **Zip** _____

Send to 3M Office Systems Division, Bldg. 216-2N, 3M Center, St. Paul, MN 55144. Attn: G. Collins.

3M

THE FISH WON'T BE BITING IN TWITCHELL CREEK TODAY,

OR TOMORROW, EITHER.



Twitchell Creek on Woods Lake in the Adirondacks. Once a great fishing spot. But no more. The fish are gone now . . . killed off by the rain. Acid rain which poisoned the water, the land, and many of the creatures who once lived here.

Acid rain is deadly because it contains nitric acid and sulfuric acid. These killers are born when rain water mixes with nitrogen oxide and sulfur dioxide . . . two chemicals which are being spewed by the ton-load into our air every day from coal-burning power plants and industrial boilers.

Some lakes contain neutralizing "buffers" which lessen acid damage. But who will stop the rain in the Adirondacks, in western Virginia, throughout New England? The Izaak Walton League, that's who. Formed in 1922 by a handful of conservation-minded fishermen, the League is now 50,000 strong. And ready to fight those who want to relax the 1970 Clean Air regulations to allow for greater sulfur dioxide emissions. We want fewer emissions. Stricter laws. And we'll get them. Because there's something at stake here far greater than fishing. And that's life itself.

For more information on our activities, write:

THE IZAAK WALTON LEAGUE OF AMERICA INCORPORATED

Izaak Walton League
1800 North Kent Street
Arlington, Virginia 22209

LETTERS

Sirs:

I read Michael McCloskey's article "Intuitive Physics" [SCIENTIFIC AMERICAN, April] with great interest because it added one more to the list of scientific disciplines that are seriously considering their counterparts in the "real" world. Intuitive physics thus takes its place among other systems such as folk medicine, folk taxonomy, folk linguistics and so forth. I found it disconcerting, however, that the concepts of intuitive physics were repeatedly referred to as misconceptions, misperceptions or mistaken beliefs. Surely correctness can be defined only within a given system of rules, not across systems (for example, whether driving on the left side of the road is correct or not depends on the set of applicable traffic regulations).

Any system of knowledge can be described as having a characteristic symbolic representation, a characteristic set of logical elements, operators and rules of inference (including a law of causality), a characteristic set of spatial and temporal coordinates and so on. The system of knowledge that includes intuitive physics and the other folk sciences shares a common symbolic representation, namely natural language, and a common logical system, namely common sense, whereas Newtonian physics and the other scientific disciplines share a common mathematical system of representation and logic. I believe, and McCloskey's article appears to confirm, that such systems can and do coexist, so that physics students, for example, can solve (Newtonian) mathematical equations for objects in motion while expressing (intuitive) descriptions of objects in motion in natural language.

The theory of coexistent systems elucidates two of the key experiments in McCloskey's study. In both cases the instructions were given in a natural language, thereby creating a coherent though non-Newtonian universe; the answers (or the reasoning leading to specific actions) were expressed in the same language and in the same universe. In particular, in this universe *to drop an object* means "to let an object fall vertically," *and* usually means "and then" (that is, "sequential" rather than "simultaneous") and *circular motion* may mean "rotation" or "spin." This last equation figures, for example, in the description of "curve balls" in baseball (and equivalent shots in other ball games); in fact, McCloskey states that "the concept of circular impetus was invoked to explain . . . the continued rotation of a wheel."

Those students who reacted in terms of intuitive physics did not (or could not) translate natural language into scientific language; they are, not surprisingly, akin to the medieval philosophers, who were essentially "natural language" philosophers. The rise of modern science is intimately connected with that of mathematics; at least since Newton the language of science and natural language have become increasingly distinct and, perhaps, mutually unintelligible. Those interested in improving science instruction may need to take a close look at the methodology of foreign language and bilingual instruction, and at the (linguistic) concept of coexistent systems.

BURCKHARD MOHR

California State University,
Dominguez Hills
Carson, Calif.

Sirs:

I found Michael McCloskey's views on intuitive physics thought-provoking. I am not convinced, however, that he is justified in implying that the pre-Newtonian ways of thinking employed by so many people are intuitive. It may be of value to look for the causes of such perceptions in learned traits. Perhaps the commonly held Aristotelian beliefs are the product of informal learning processes that inspire a confusion of frames of reference and the acquisition of misconceptions of such concepts as force, mass and velocity. Perhaps such erroneous views persist because they are handed down to the young by ill-advised parents and elementary school teachers. Even before they can talk the young are subjected to cartoon characters and things that operate in a world where the laws of motion are subject to the arbitrary whims of the artist's brush.

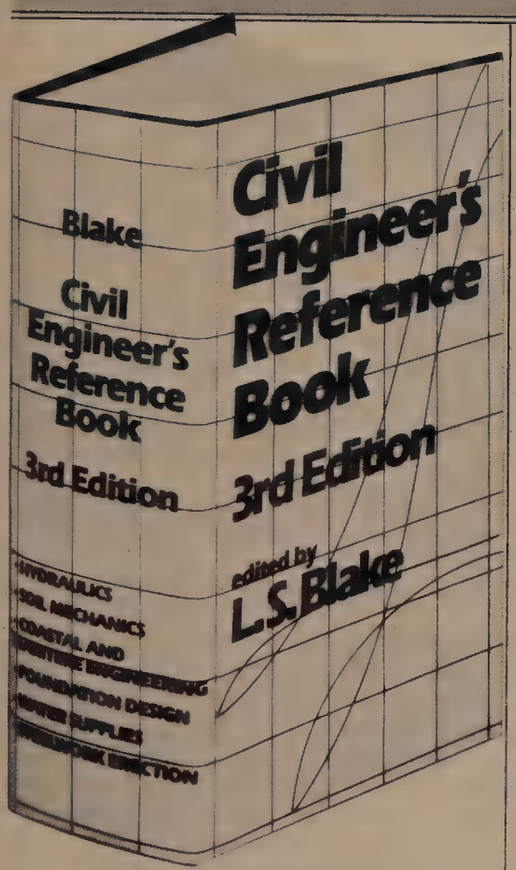
It appears that many people by high school age have constructed a conceptual framework, albeit a mistaken one, with which they are at ease and that allows them to deal with their world to their satisfaction. I should like to have seen some record of the reasons people opt for one explanation or description of motions rather than another. That some people who have had no formal exposure to Newton hold mistaken beliefs about motion does not surprise me. It is disquieting, however, that such a high proportion of those who have come through an elementary dynamics course cling firmly to pre-Newtonian conceptions of motion.

As one who is involved in teaching Newtonian physics to high school students I am struck by the lack of imagination in methods of presenting an efficacious path to the classical view. Most high school texts and programs lead the student from Newton's first law, through his second to the third. The Harvard Physics Project, the Physical

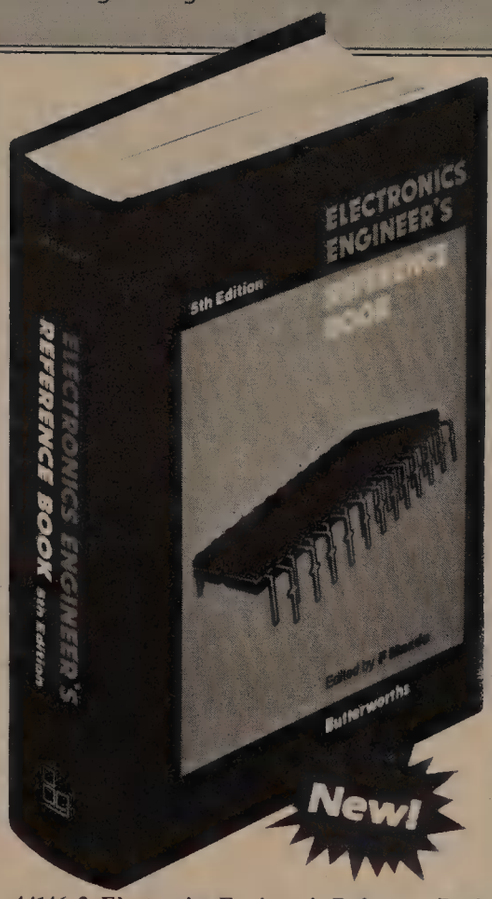
Take 1 of three Engineer's Reference Books for only $2.95

when you join any one of our three Professional Engineering Book Clubs—*the Electrical and Electronics Engineering Book Club, the Professional Civil Engineering Book Club* or the *Professional Mechanical Engineering Book Club.* Whichever club you join, you simply agree to buy 3 more books—at handsome discounts—within the next 12 months.

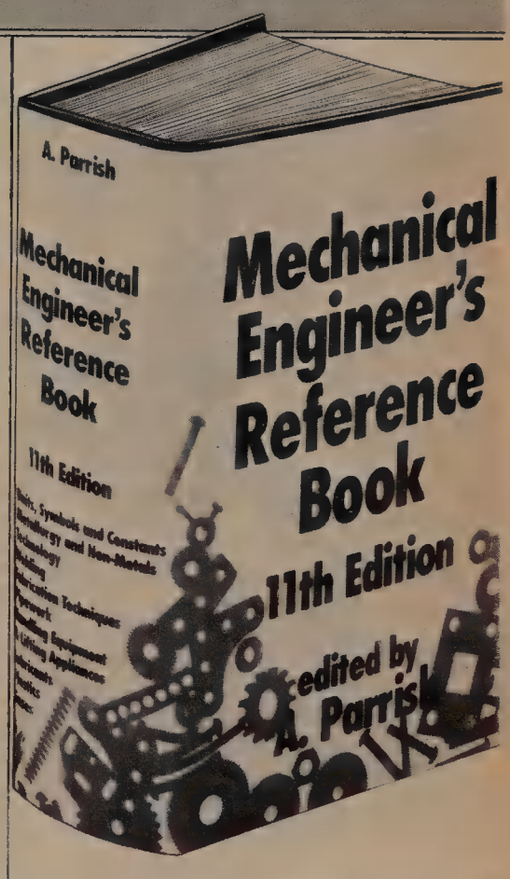
Each of the Professional Engineering Book Clubs provides a convenient and inexpensive way to improve your professional skills. The books we offer keep you up-to-date about new concepts, techniques and systems which are constantly being developed. Whichever field in engineering is your specialty—Electrical and Electronics, Civil or Mechanical—you can be sure that we have the ideas and answers to help you stay on top of your profession. So take the Engineer's Reference Book of your choice and join one of our three clubs.



38250-3. Civil Engineer's Reference Book. *Edited by L.S. Blake.* Latest edition of this engineering classic. Concise chapters on mathematics and statistics, site investigation, hydraulics and surveying, soil and rock mechanics, sewage disposal, contract management and insurance. Current practices in design, structures and use of materials. Thousands of diagrams, charts, formulas, graphs and illustrations. 1800 pages. Publisher's Price: $84.00.



44146-3. Electronics Engineer's Reference Book. Fifth Edition.
Edited by F. Mazda.
Extensively rearranged and revised text. With a new format and more conventional size volume. The emphasis is on the latest technologies with 32 *new* chapters on topics like semiconductor memories, microprocessors, fibre-optic communications, and office communication. The 62 total chapters cover: techniques, physical phenomena, materials and components, electronic design, and application areas. Extensively illustrated and indexed. Publisher's Price: $119.00.



61655-3. Mechanical Engineer's Reference Book. *Edited by A. Parrish.* Revised and updated. Comprehensive description of theory and design data and fabrication of materials. Chapters on thermodynamics, fluid mechanics, metallurgy, lubricants, computers, productivity measurement, instrumentation and much more. Extensive symbols and abbreviations. Thousands of charts, diagrams, formulas and illustrations. 20 sections contributed by 37 engineering consultants. 1700 pages. Publisher's Price: $79.95.

4 Good Reasons to Join

1. The Finest Books. Of the hundreds and hundreds of books submitted to us each year, only the very finest are selected and offered. Moreover, our books are always of equal quality to publishers' editions, *never* economy editions.

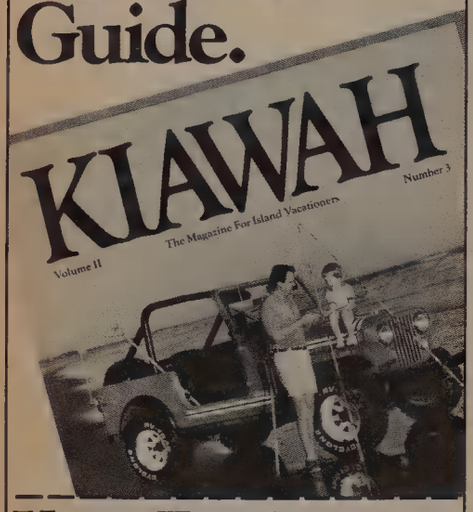
2. Big Savings. In addition to getting the Engineer's Reference Book of your choice for only $2.95 when you join, you keep saving substantially—up to 30% and occasionally even more. (For example, your total savings as a trial member—including this introductory offer—can easily be over 50%. That's like getting every other book free!)

3. Bonus Books. Also, you will immediately become eligible to participate in our Bonus Book Plan, with savings of up to 70% off the publishers' prices.

4. Convenient Service. At 3-4 week intervals (16 times per year) you will receive the Book Club News, describing the Main Selection and Alternate Selections, together with a dated reply card. If you want the Main Selection, do nothing and it will be sent to you automatically. If you prefer another selection, or no book at all, simply indicate your choice on the card, and return it by the date specified. You will have at least 10 days to decide. If, because of late mail delivery of the News, you should receive a book you do not want, we guarantee return postage.

If reply card has been removed, please write to: **Professional Engineering Book Clubs, Dept. —-AJ5 Riverside, N.J. 08075** to obtain membership information and an application.

Colorful Fall Vacation Guide.



Yours Free!

Send for the 1983 edition of our full-color 30-page magazine, because this year Kiawah offers more island than ever. Enjoy twice the golf, twice the tennis and a superb new restaurant. Then discover our 10 miles of beach, exciting Jeep Safaris, fascinating shops and all the charm of nearby historic Charleston. For reservations, call us, or your travel agent.

Name_____

Address_____

City/State/Zip_____

Kiawah Island

Dept. VMS12
P.O. Box 12910
Charleston, SC 29412
1-803-768-2121

Call TOLL FREE 1-800-845-2471.



Better Than Jogging, Swimming or Cycling

NordicTrack

Jarless Total Body Cardiovascular Exerciser

Duplicates X-C Skiing For The Best Way To Fitness

NordicTrack duplicates the smooth rhythmic total body motion of XC Skiing. Recognized by health authorities as the most effective fitness building exercise available. Uniformly exercises more muscles than jogging, swimming, cycling and rowing.

Does Not cause joint or back problems as in jogging. Highly effective for weight control and muscle toning.

Easily Adjustable for arm resistance, leg resistance and body height. Smooth, quiet action. Folds compactly to require only 15 by 17 inches of storage area. Lifetime quality.

Used In thousands of homes and many major health clubs, universities, and corporate fitness centers.

Call or Write for FREE BROCHURE

Toll Free 1-800-328-5888 MN 612-448-6987

PSI 124 F Columbia Crt., Chaska, Minn. 55318

Sciences Study Committee and the Nuffield curriculum are notable among the guilty. There operates the implicit belief that the shortest distance between two points is always a straight line.

In the reality of the high school physics laboratory I have found that conceptual development in the classical mode can be achieved with high success by considering the second law first and regarding the first as its special case. Students can weave the third law into their conceptual fabric by perceiving it, as it is commonly stated, as an expression of the principle of conservation of momentum. There are probably several successful routes to Newton. I doubt whether the predominating Western educational paradigm qualifies as one. Perhaps the time has come to address the mind-set that denies to so many an access to the Newtonian perspective.

F. McGeachy

Henry Wise Wood High School
Calgary, Alberta

Sirs:

Michael McCloskey's article should be required reading for football officials. Many seem to think that a ball thrown perpendicular to the direction of the thrower's motion will not advance and hence is not a forward pass. An example of such loose interpretation occurred during last year's (in)famous five-lateral Rugby Runback in the Stanford-California Big Game. Although the final over-the-shoulder toss was directed backward, the films show a net ball advance of two yards. This should not surprise anyone who stops to think about it, but then not many people there were capable of higher mental functioning after that play.

Thomas Tollefsen

Glen Ellen, Calif.

Sirs:

These letters raise a number of interesting points. Professor Mohr suggests that the intuitive beliefs I described should not be referred to as misconceptions because correctness can be defined only within a given set of rules. As an example he mentions that driving on the left side of the road may be correct or incorrect, depending on what set of traffic rules is applicable. This argument misses an important point. For both driving behavior and beliefs about motion correctness is defined by reference to some standard. For driving the standard is the applicable traffic laws. For beliefs about motion the standard is the actual behavior of moving objects. The intuitive beliefs I discussed are misconceptions because they are inconsistent with the actual behavior of objects. For example, the belief that an object dropped by a walking person will fall in a straight vertical trajectory is a misconception because such an object will move forward as it falls.

Professor Mohr also seems to overstate the connection between natural language and intuitive science. He appears to think natural language is inextricably tied to intuitive physics as opposed to Newtonian physics, so that instructions given in natural language create "a coherent though non-Newtonian universe," and application of Newtonian principles requires a translation from natural to scientific language. It may well be the case that natural language to some extent reflects intuitive beliefs about the world, so that principles of intuitive physics are more easily expressible in natural language than the principles of Newtonian physics. It is by no means the case, however, that the use of natural language forces the adoption of a non-Newtonian framework. Newtonian principles as well as intuitive ones can be expressed in natural language (as I hope the discussion of Newtonian mechanics in my article illustrates).

McGeachy points out that an important issue in the study of intuitive physics is: How do people come to develop the intuitive beliefs? In the article I suggested that at least some beliefs may stem from illusions in the perception of motion. McGeachy suggests that cartoons may also be responsible in part. It is certainly true that objects in cartoons frequently behave in a non-Newtonian manner. A striking example is the coyote in the Roadrunner animated cartoons who, after racing over the edge of a cliff, continues horizontally for some time, comes to a stop and then falls straight down. As I mentioned in the article, a small percentage of people believe that an object going over the edge of a cliff will in fact behave this way. Thus cartoons may well contribute to the development of non-Newtonian intuitive beliefs about motion.

Mr. Tollefsen makes an interesting suggestion about the controversial ruling by officials in last year's Stanford-California football game. It is a common misconception that the trajectory of an object thrown by a moving person will not be affected by the person's motion. Like the belief that an object carried by a moving person falls straight down when it is dropped, this misconception is a manifestation of the idea that a carried object does not acquire impetus from the carrier.

Michael McCloskey

Johns Hopkins University
Baltimore, Md.



Once a year in Springtime, there occurs an event in Asia that sweater wearers the world over might celebrate: *the moulting of the Kashmir goat.*

This event provides the annual supply of cashmere, that rare and wondrous hair of the goat that is to the world of natural fibers what gold is to the supply of baser metals.

Cashmere—the downy undergrowth produced by the Kashmir goat—a material so rare it rewards a climb of the Himalayas. A material we take pains to travel around the world unchanged, except for a C instead of a K when it reaches the Scottish knitting mills which Lands' End haunts to process it. And to make some of the world's greatest sweaters.

More like wool than any fiber.
Natural cashmere is brownish in color. It yields itself to the spinning of luxurious, soft, wool-like yarns with a characteristic highly napped finish.

It is highly prized—not least by us at Lands' End and our customers—for its softness, and warmth. It is lighter in weight than wool, but not as durable. As they say, nobody is perfect—not even the Kashmir goat.

Our cashmeres are rare values.
Lands' End specifies that our cashmere sweaters be knitted from two-ply yarn, which is extremely fine and close and extremely soft. By the same token it is a costly material—there being a strict limitation on the total amount of cashmere hair available.

Still, we have managed to lay hold on an encouraging assortment of cashmere sweaters and we've priced them affordably.

(We don't haunt those Scottish mills for nothing.)

They include cashmere cable crew-necks with hand-looped collars, as well as Fair Isle styles for her, sleeveless V-necks for him, and long-sleeved V-necks for the two of them.

Seeing is believing and thrilling.
Obviously you should see our cashmeres on people and in our catalog—and what could be easier than to get a free catalog simply by dialing our toll-free number (800-356-4444) 24 hours a day. Or sending in the coupon below.

Remember, like everything else in the catalog—from shoes to soft luggage to summer wear and accessories—your cashmere sweater is backed by our short, sweet, two word guarantee:

GUARANTEED. PERIOD.

There is no fine print.

Call now. So next year when the Kashmir goat performs his miracle, you can celebrate in Lands' End style.

LANDS' END
DIRECT MERCHANTS

of fine wool and cotton sweaters, Oxford button-down shirts, traditional dress clothing, snow wear, deck wear, original Lands' End soft luggage and a multitude of other quality goods from around the world.

☐ **Please send free catalog.**



Lands' End Dept. Q-25
Dodgeville, WI 53595

Name_____

Address_____

City_____

State_____ Zip_____

Or call Toll-free:

800-356-4444

(Except Alaska and Hawaii call 608-935-2788)

50 AND 100 YEARS AGO

SCIENTIFIC AMERICAN

SEPTEMBER, 1933: "A secret method of depositing photo-sensitive elements on a four-inch square of mica is the foundation of a new television transmission system recently announced by Dr. Vladimir K. Zworykin of the RCA-Victor Research Laboratories. Without mechanically moving parts, Dr. Zworykin has succeeded in transmitting images built up 250 lines to the inch, or 62,500 elements to the square inch. The best that can be done with the conventional scanning disk is about 50 lines or 2,500 picture elements. The modified cathode-ray tube used in the Zworykin transmitting system is called the Iconoscope. The mica mosaic is in such a position that an image can be focused upon it through the glass of the tube. At the base of the tube is an electron-producing 'gun.' The invisible stream of projectiles from the gun is controlled by four coils mounted on a frame and placed outside the tube. When the proper alternating currents from vacuum-tube oscillators are applied to the coils, the beam of electrons can be controlled so that it will sweep back and forth over the surface of the mosaic, covering every portion of it 24 times during each second of operation. The mosaic plate has on one surface 3,000,000 light-sensitive units and on the other surface a plate of silver. Each of the units, together with this plate, forms a tiny condenser that stores up current in exact proportion to the amount of light reaching the unit. This condenser is discharged as the electron beam sweeps over it, and therefore adds its bit of current to the 'picture signal' being built up for transmission. Dr. Zworykin states that this new system will operate at a speed comparable to that of motion-picture camera film. Thus it is possible to televise objects in ordinary light and to dispense with the intense lighting systems that up to now have been necessary. At the receiving end the Zworykin cathode-ray type of reproducer is used."

"Our spectroscopes reveal multitudes of lines in the spectra of celestial bodies. For almost all the metals these lines are in the visible or near ultra-violet region and accessible to observation, hence our test for metals in the stars is a delicate one. But for the lighter non-metallic atoms these strong lines lie far out in the ultra-violet. Here we come to the most tantalizing thing imaginable, at least to the astrophysicist. The earth's atmosphere, which is sometimes fairly transparent to visible light and to a smallish region in the ultra-violet adjacent to the visible, is as opaque as a stone wall to the shorter waves. If only we could observe the far ultra-violet part of the spectra, we would doubtless find lines enormously stronger than the rather feeble ones that appear on our photographs. The stellar spectra we photograph are mere wraiths of the far more remarkable ones we could observe if we could find some way of moving and keeping alive on an airless body such as the moon. The good spectroscopist might perhaps be allowed to go, instruments and all, to set up an observatory on the moon when he died."

"Last year United Air Lines ordered 60 specially built airplanes from the Boeing Airplane Company at Seattle. The characteristics of these new-type planes reveal the lessons the air-transport company has learned from 50,000,000 miles of flying with airmail, passengers and express. The airplane is a low-wing metal monoplane with two Wasp engines supercharged to deliver 550 horsepower each at 5,000 feet above sea level. It is fast, having a top speed of 175 miles per hour and a cruising speed of 155 miles per hour with full load. It is medium in size, accommodating 10 passengers with their baggage, two pilots and a load of better than a third of a ton of cargo. It is comfortable, its roomy passenger cabin being equipped with chairs larger than any previously used in commercial transport planes, and spaced 40 inches apart to provide that all-desirable leg-room for passengers."



SEPTEMBER, 1883: "The chief questions before the annual meeting of the American Association for the Advancement of Science in Minneapolis were undoubtedly two, viz., on the theory of evolution and on glacial action. President Dawson initiated the discussion in his retiring address. His utterances were judicious and respectful in their tone but so decidedly in opposition to the extreme evolutionists as to kindle excitement and provoke replies. On the grand question of glacial action, Dr. Dawson took ground against the origin of the glacial drift in a great continental glacier, claiming that there was instead a wide glacial sea with Arctic currents and icebergs with here and there local glaciers. The gauntlet thus thrown down was lifted by those adhering to the notion of a continental glacier. The geological room became too crowded for comfort, and the closing discussions were transferred to the large chapel of the university."

"Our readers are well acquainted with the herculean struggle of the British anti-vaccinationists, now continued for 10 or 15 years, directed against the compulsory vaccination law. From year to year the movement has been gaining strength. Societies have been organized in all quarters, periodicals established, large funds contributed and many of the leading men and women of England, including not a few of the nobility, enlisted in the enterprise. Judging from the clamor that filled the air, the heart and head of the kingdom were gained over to the enterprise. At last the long labor of the mountain culminated in the introduction of a repealing bill in the House of Commons. The hour of promise was come, and a triumphant majority of the representatives of the people of Great Britain would reward the labors and verify the sanguine predictions of the anti-vaccinationists. The vote was taken; of 302 members present, 16 voted for repeal and 286 against!"

"General George B. McClellan, who has recently visited many parts of the Texas Panhandle, predicts that by the year 1890 the state will have a population of 5,000,000, and he also affirms it can support 20,000,000 without overcrowding. The capabilities of Texas are only just being discovered; it is larger than France, with a better soil and an equable climate, is well watered and is being completely intersected by railroads. There was a large increase of population between 1870 and 1880, and there will be a still larger one in the present decade. The state is already second only to Georgia in the production of cotton, and it produces more cattle than any other two states."

"Several learned professors in Ohio educational institutions having recently embarked in a newspaper controversy as to whether a skillful base ball pitcher could or could not throw a ball in a horizontal curve, the question was set to rest by actual experiment. A straight chalk line was drawn between the home plate and the pitcher's position. Next to the right-handed pitcher's position a board was placed so that the pitcher could not cross the line. Half way between the pitcher's position and the home plate a section of picket fence was set on the line, and close to the home plate a second section was placed on the line but on the opposite side. The ball was delivered from one side of the line with a right-hand twist. It passed to the right of the first section of picket fence and to the left of the second section, striking the ground on the same side of the line as the pitcher."


Imported by Rémy Martin Amerique, Inc., N.Y., N.Y.
Sole U.S.A. Distributor, Première Wine Merchants Inc., N.Y., N.Y. 80 Proof.
Savor the
sense of Rémy.
REMY MARTIN
V.S.O.P.
COGNAC
Rémy
REMY MARTIN V.S.O.P. COGNAC. SINCE 1724.

How to turn data you can see.

There are more people in more places doing more things with Apples than with any other personal computer in the world.

Which is saying a lot.

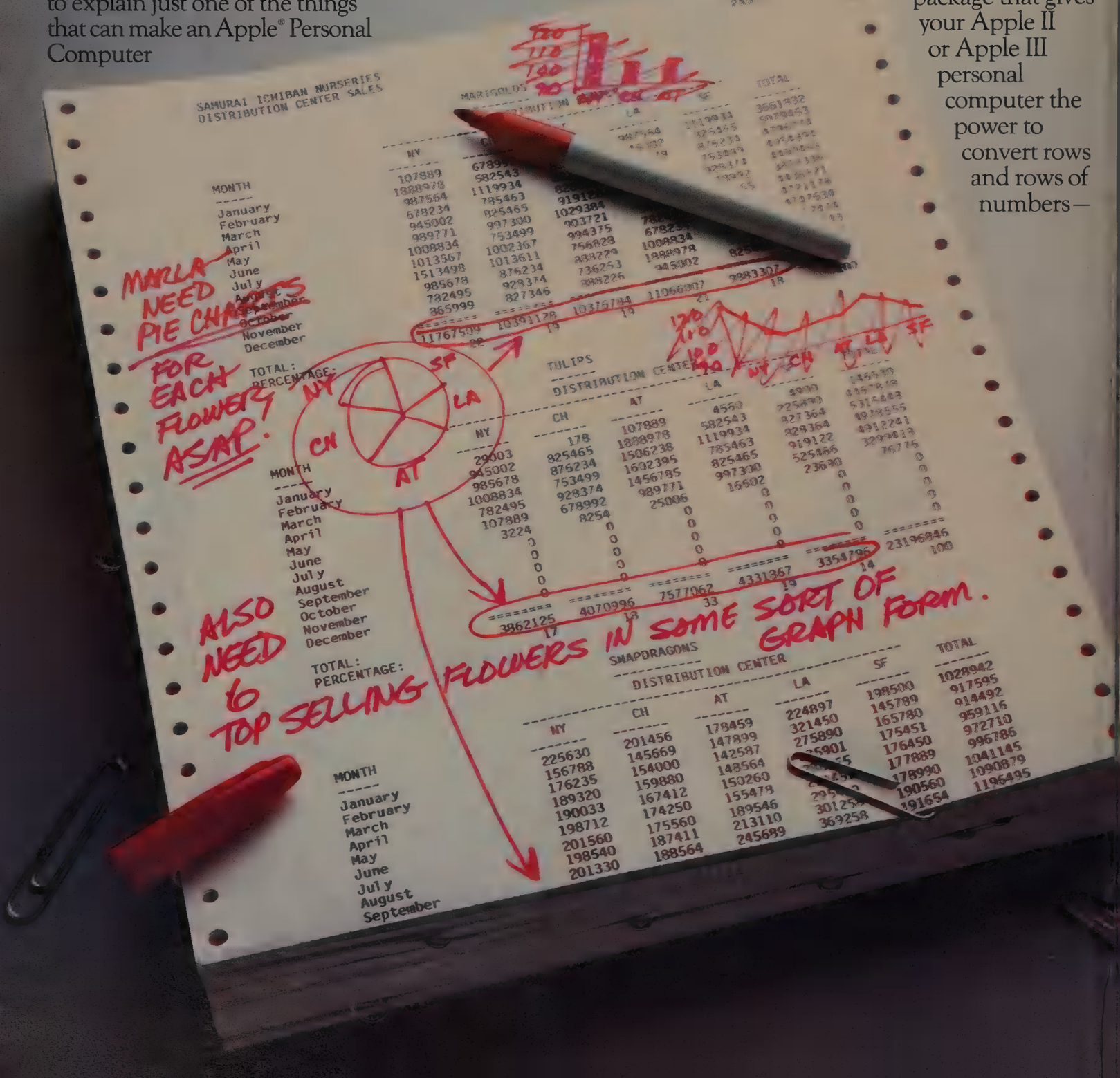
But we'd like to take the time to explain just one of the things that can make an Apple® Personal Computer more meaningful to you, personally.

It's called business graphics.

An important business tool that demonstrates, quite graphically, how one picture is worth a few thousand numbers.

More ways to see what you're doing.

That's the beauty of Apple Business Graphics. The software package that gives your Apple II or Apple III personal computer the power to convert rows and rows of numbers—



a sea of data into

or pounds of printout—into one simple, colorful, comprehensible illustration.

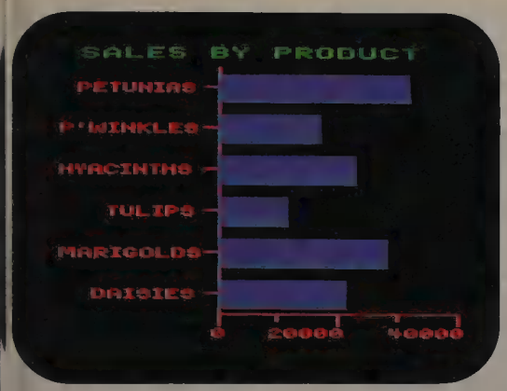
And that gives you the power, for example, to thoroughly understand relationships and comparisons in a complex market analysis. Without having to spend hours sifting through dull data—armed with little more than a ruler, a compass and a few not-so-magic markers.

With Apple Business Graphics, all you do is enter the data or transfer it directly from your VisiCalc® model or a number of other computer files. Your Apple will take it from there.

Helping you to get the big picture in a variety of ways. Line or area graphs, bar graphs, pie charts or scattergrams.

Whichever configuration best suits your application.

When you're through charting market shares you can do the same with forecasts, budgets, stock trends, business plans or customer demographics.



Horizontal bar graph (a unique feature) makes it easy to compare the profitability of six different products – from periwinkles to petunias.

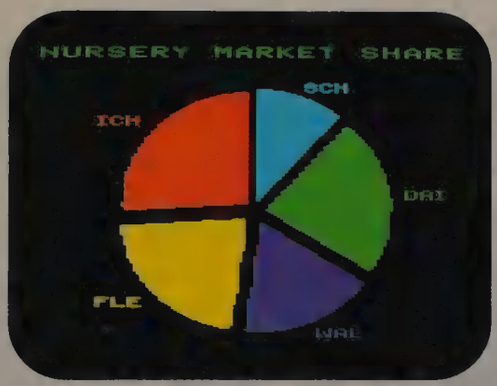
Show your stuff.

In addition to helping you better visualize results, Apple Business Graphics can make presentations more presentable.

Because whatever you've produced, including charts and graphs, can be rapidly, easily and colorfully reproduced on a printer or plotter.

Moreover, sophisticated Apple features like exploded views, unlimited overlays and floating titles allow your displays to be remarkably flexible.

By combining Apple Business Graphics with various slide-show packages, charts and graphs on your monitor can easily be converted into impressive slides and transparencies.



To get the overall market share picture, (and any other percentage related data) Ichiban management consults this easy-to-digest pie chart.



Getting started in pictures.

Start by seeing any of the over 1300 authorized Apple dealers.

They'll show you Apple Business Graphics software can generate more types of pictures, in more colors, using more data than any other graphics package.

More important, they'll show you what all that can mean to you, your accounting firm, your import business, or your chain of quiche-on-a-bun family restaurants.

They'll even help you produce your first picture.

But business graphics is just one of the things an Apple can help you do. One of thousands. Which is not terribly surprising, when you consider there's more software available for Apples than for any other personal computer.

So consider a trip to your Apple dealer. And draw your own conclusions.



With up to six colors per graph, employees can easily see just how their annual sales blossomed relative to the rest of the market.

apple

Call (800) 538-9696 for the location of the authorized Apple dealer nearest you, or for information regarding our National Account Program.
In Canada, call (800) 268-7796 or (800) 268-7637. Or write Apple Computer Inc., Advertising and Promotion Dept., 20525 Mariani Ave., Cupertino, CA 95014.
Visicalc is a registered trademark of VisiCorp. © 1983 Apple Computer Inc.

SCIENCE/SCOPE

A Very High Speed Integrated Circuit chip built for the U.S. military uses technology that makes it inherently hardened against radiation. The chip, produced after less than two years of development, draws on complementary metal oxide semiconductor/silicon on sapphire technology. It has circuit dimensions of 1.25 micrometers, or about 50 millionths of an inch. The VHSIC program is being conducted by the Department of Defense to develop chips that will give electronic systems a tenfold increase in signal processing capability. The high-speed, compact VHSIC chips will be more reliable and will require less power than integrated circuits now in use. Hughes Aircraft Company is the only contractor in the tri-service program pursuing CMOS/SOS technology.

The new AMRAAM missile will be good at evading enemy detection through a clever improvement to its radar system. The improvement, now patent pending, is done simply and with only a little extra hardware. It greatly reduces inaccuracies caused when the missile jumps from one radar frequency to another en route to its target. Frequency hopping makes it extremely difficult for enemy radar-detection equipment to get a fix on the missile. Hughes designed and developed the Advanced Medium-Range Air-to-Air Missile for the U.S. Air Force and Navy.

Of the improvements in productivity of electronics offered by computers, some of the most dramatic can be found on the manufacturing floor. Computer-controlled automation yields important savings through increased efficiency, flexibility, and accuracy. Computers can repeat virtually all processes -- machining, chemical processing, fabrication and assembly, quality inspection, and testing -- with infallible precision. In the production of digital electronics modules at Hughes, productivity sometimes has been increased by a factor of 10 or more. Hughes is spending $240 million over five years on computer-aided manufacturing.

The two shortwave infrared bands on Landsat 4's thematic mapper are gathering data that sensors on previous Earth resources satellites couldn't. These bands, which are sensitive to the amount of water in plant leaves, will identify plants and assess their health. They can map snow cover without being fooled by clouds because snow appears very dark, while clouds remain bright. The infrared bands also detect a wider variety of rock and soil types. Experimental studies showed these bands can identify variations in type and abundance of clay minerals exposed at the Earth's surface. This information can be used to substantially improve the quality of geological maps. Hughes and its Santa Barbara Research Center subsidiary built the thematic mapper for NASA.

Career growth opportunities exist at all levels at Hughes Support Systems for a variety of engineers qualified by degree or extensive work experience. They include systems engineers and software and hardware design engineers for major simulation and test equipment programs. Also, field engineering posts throughout the U.S. offer travel, autonomy, and responsibility for the life cycle of Hughes electronics systems. Please send your resume to Lowell Anderson, Professional Employment, Dept. SE, Hughes Aircraft Company, P.O. Box 9399, Long Beach, CA 90801-0463. Equal opportunity employer.

Creating a new world with electronics

HUGHES

HUGHES AIRCRAFT COMPANY

For more information please write:
P.O. Box 11803, Los Angeles, CA 90291

THE AUTHORS

RAYMOND SIEVER ("The Dynamic Earth") is professor of geology at Harvard University. He has three degrees from the University of Chicago: a B.S. (1943), an M.S. (1947) and a Ph.D. (1950). While he was still a graduate student he joined the Illinois Geological Survey, for which he worked until 1957. In 1956 he became a member of the Harvard faculty, and he has remained at Harvard since. He has served two stints as chairman of the department of geology. He recently stepped down for the second time to devote himself to teaching and experimental work. Siever writes: "My current interests are in the extremely low rates of geochemical reactions in buried sediments. I am fascinated by the relation of these processes to the internal heat of the earth, that is, how the formation of sedimentary rocks near the surface is affected by the dynamics of the interior, hundreds of kilometers below. Away from the university, the piano remains my abiding interest: alone, in duets or in chamber music."

RAYMOND JEANLOZ ("The Earth's Core") is associate professor of geology and geophysics at the University of California at Berkeley. He was graduated from Amherst College with a B.A. in 1975 and went on to obtain his Ph.D. in geology and geophysics at the California Institute of Technology in 1979. Jeanloz spent two years as a member of the faculty at Harvard University before moving to Berkeley in 1981. He writes: "My main interest is in mineral physics. I am particularly interested in studying how the physical and chemical properties of minerals control the ways in which planets evolve. Much of the work involves studying materials at the high pressures and temperatures that occur deep inside the planets. Like many earth scientists I became interested in geology through my enjoyment of the outdoors, in particular the high mountains."

D. P. McKENZIE ("The Earth's Mantle") is reader in tectonics at the University of Cambridge. His association with Cambridge goes back to his student days. He became a graduate student of Sir Edward Bullard's in 1963 and got his Ph.D. in geophysics at Cambridge in 1966. Soon after receiving his degree he joined the Cambridge faculty. He was promoted to his present job in 1979. He has served as visiting professor at several American institutions, most recently the University of Chicago. Most of McKenzie's scientific work has been on plate tectonics; he is currently most interested in the process of convection in the mantle and in the evolution of sedimentary basins.

JEAN FRANCHETEAU ("The Oceanic Crust") is a physicist on the staff of the Institut de Physique du Globe of the University of Paris. He is a native of France who received his diploma in mining engineering at the École des Mines in Nancy. He came to the U.S. to continue his education; the Scripps Institution of Oceanography awarded him a Ph.D. in geophysics in 1970. After obtaining his Ph.D. he returned to France to join the staff of the Centre Océanologique de Bretagne in Brest. He left the center in 1981 to go to the University of Paris. Francheteau's main scientific interest is the exploration of the ocean floor. His work has concerned the structure and morphology of the oceanic crust and the tectonics and geophysics of the ocean bottom. Much of his recent work has been done on the sea floor in the French submersible *Cyana*. He was the leader of Project RITA in 1978. During the expedition the first dives in a manned submersible to the crest of the East Pacific Rise were made in the *Cyana*.

B. CLARK BURCHFIEL ("The Continental Crust") is professor of geology at the Massachusetts Institute of Technology. His B.S., earned in 1957, is from Stanford University. He went on to get an M.S. at Stanford in 1958 before getting his Ph.D. in geology from Yale University in 1961. In the same year he joined the faculty at Rice University; he was at Rice until 1976. At the end of that time he moved to M.I.T. Burchfiel writes: "My current interest is in the process of orogenesis (mountain building) and its relation to the interactions of tectonic plates and to intraplate deformation. The specific areas I study are western North America, the Alpine mountain chains of eastern Europe and (more recently) of north-central China."

WALLACE S. BROECKER ("The Ocean") is professor of geology and director of the laboratory of geochemistry at Columbia University. His association with Columbia is more than three decades old. He received three academic degrees there: a B.A. in 1953, an M.A. in 1956 and a Ph.D. in 1958. In 1956, while he was still working on his Ph.D., he became a member of the Columbia faculty, and he has remained there since. Broecker's scientific concerns are diverse. He has done work on ancient climates and the causes of the cycles of glaciation on the earth, the large-scale mixing of water in the oceans and environmental geochemistry, including the effects of acid rain, metals in lakes and ocean chemistry.

ANDREW P. INGERSOLL ("The Atmosphere") is professor of planetary science at the California Institute of Technology. He attended Amherst College as an undergraduate, getting his B.A. in 1960. He received his postgraduate education at Harvard University, which awarded him an M.A. in 1961 and a Ph.D. in atmospheric physics in 1965. He writes: "My career goals took shape while I was an undergraduate physics major and a summer research assistant at the Woods Hole Oceanographic Institution. Since getting my Ph.D. I have devoted my professional life to the study of atmospheres and climates. I have participated in several National Aeronautics and Space Administration missions to the other planets, including the Pioneer missions to Jupiter and Saturn, the Pioneer mission to Venus and the Voyager mission, which flew by Jupiter and Saturn in 1979–81 and will fly by Uranus in 1986. I am also a member of a team that is analyzing the earth's radiation budget by means of data from a Nimbus spacecraft."

PRESTON CLOUD ("The Biosphere") is professor emeritus of biogeology and environmental studies at the University of California at Santa Barbara. He got a B.S. at George Washington University in 1938 and a Ph.D. from Yale University in 1940. From 1942 to 1961 he was on the staff of the U.S. Geological Survey. During his years with the Survey he held several different positions, among them that of chief of paleontology and stratigraphy. When he left, he became a member of the faculty at the University of Minnesota and the University of California at Los Angeles before moving to Santa Barbara in 1968. Cloud's main interest is the interaction of the physical environment and life in early times. He writes: "My interest in the origin of multicellular animal life, its antecedents and its first appearance on the earth has long been the central aspect of my research. In 1948 I proposed that multicellular animal life first appeared and evolved into a diversity of forms during an interval of less than 100 million years at the beginning of Phanerozoic time, a view that is now widely accepted. Using stromatolites, sedimentary structures, microbial fossils and the methods of ultramicroscopy and geochemistry, I traced the record of microbial life far behind the dawn of the Cambrian and integrated the study of the earliest biospheric, atmospheric and chemospheric evolutions and their interactions on the primitive earth."

Instead of owning a car maybe you need a cam



In a world full of cookie-cutter SLRs, Nikon introduces the world's

Over the years millions of people have bought millions of SLRs designed for beginners.

But what happens when you don't want to take pictures like a beginner anymore?

Many stick with their mass-produced SLR and wonder why they can't get better results.

For the few, however, who are determined to improve as photographers, Nikon has created an engineering marvel called the FE2. An automatic SLR so different, there are no other cameras to compare it with.

For instance, take the matter of Selective Focusing. Which is the ability to vary apertures to express your impression of the subject. An absolute must for anyone who wants to improve as a photographer.

The FE2 gives you more options for Selective Focusing than any other 35mm SLR in the world.

To begin with, the FE2 is an aperture-priority automatic. So you have total control over depth of field when you shoot in automatic.

But, as we said, that's just the beginning.

The FE2 is the first 35mm automatic with a 1/4000 shutter speed. An innovation made possible by Nikon's development of an ultra-thin honeycomb titanium shutter. It means photographers like yourself will, for the first time, be able to use a full range of apertures in bright sunlight. And because the FE2 is faster than any other 35mm automatic, it gives you greater control over exposure than other cameras when using the

If you love photography, don't miss Nikon's "The Photographer's Eye" on SPN Cable TV. Check local listings. For further information write Dept. 50, Nikon Inc., 623 Stewart Ave., Garden City, N.Y. 11530. © Nikon Inc. 1983.

ra designed for millions
ra engineered for a few.



astest 35mm SLR automatic. The FE2.

test 1000 ASA color print film.

The FE2 is also the only 35mm utomatic with a 1/250 flash synch peed. Which means you'll have the ility to select a wider range of lens pertures when taking flash pictures utdoors. An achievement that was nce considered impossible with a 5mm SLR.

In addition, the FE2 is one of the w cameras equipped with through- ae-lens flash metering (TTL). Which ives you unparalleled accuracy ecause flash exposures are measured by the camera as they pass through the lens. And not by a sensor on the flash unit itself. TTL allows you to achieve virtually unlimited aperture settings when shooting flash indoors with a dedicated Nikon Speedlight.

Even when you're not taking photographs with the FE2 it's an astonishing camera. Its balance and 'feel' make other SLRs seem bulky and unresponsive. And its copper silumin aluminum body makes it one of the most rugged Nikons ever made.

So while other cameras are designed for millions, the FE2 is engineered to take the kind of pictures only a few strive for.

One-in-a-million.

Nikon

We take the world's greatest pictures.™

MATHEMATICAL GAMES

The topology of knots, plus the results of Douglas Hofstadter's Luring Lottery

by Martin Gardner

EDITOR'S NOTE: This is the second of two articles by Martin Gardner.

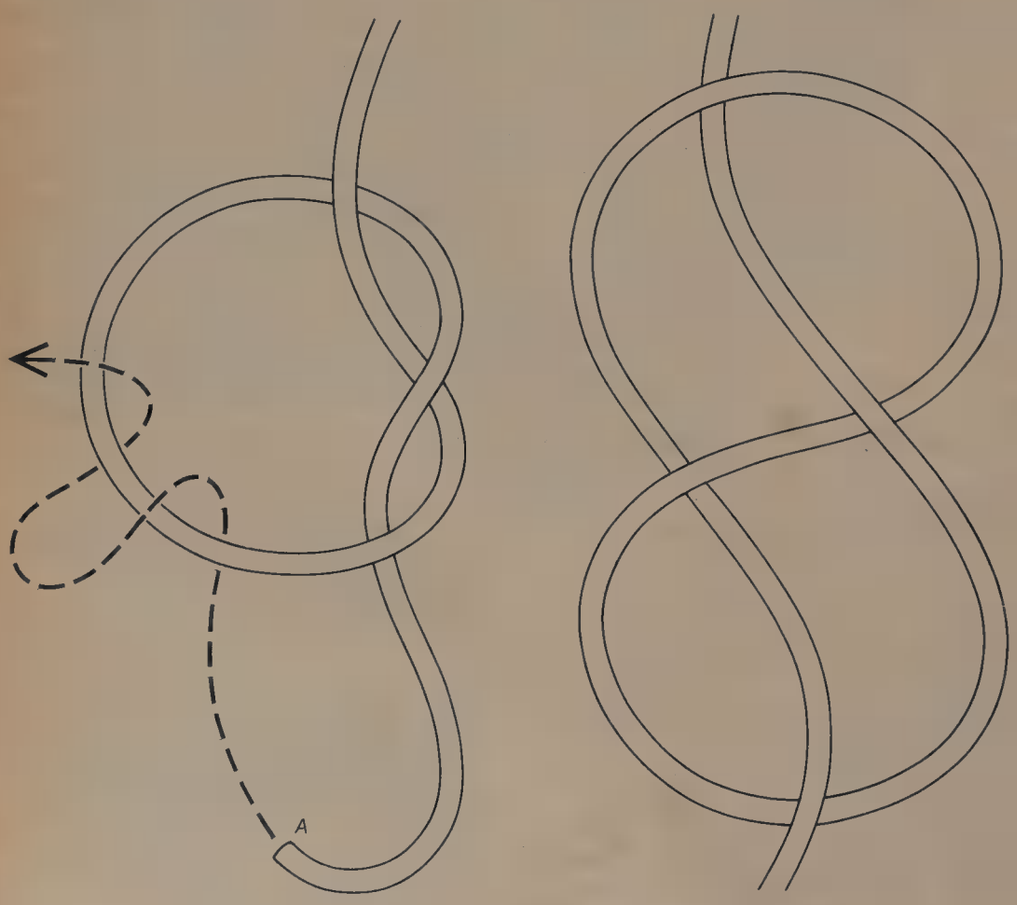
Mathematics never proves anything about anything except mathematics, and a piece of rope is a physical object and not a mathematical one. So before worrying about proofs, we must have a mathematical definition of what a knot is and another mathematical definition of when two knots are to be considered the same.

—RICHARD H. CROWELL and RALPH H. FOX, *Introduction to Knot Theory*

This is knot season in *Scientific American*. Last month in "The Amateur Scientist" Jearl Walker showed how a simple mathematical analysis could be applied to determine whether a given knot will hold or slip. This month I shall take up the topology of knots.

To a topologist knots are closed curves embedded in three-dimensional space. It is useful to model them with rope or cord and to diagram them as projections on a plane. If it is possible to manipulate a closed curve—of course, it must not be allowed to pass through itself—so that it can be projected on a plane as a curve with no crossing points, then the knot is called trivial. In ordinary discourse one would say the curve is not knotted. "Links" are two or more closed curves that cannot be separated without passing one through another.

The study of knots and links is now a flourishing branch of topology that interlocks with algebra, geometry, group theory, matrix theory, number theory and other branches of mathematics. Some idea of its depth and richness can be had from reading Lee Neuwirth's excellent article "The Theory of Knots" in *Scientific American* (June, 1979). Here we shall be concerned only with some recreational aspects of knot theory: puzzles and curiosities that to be understood require no more than the most elementary knowledge of the topic.



A trefoil knot is dissolved **(left)** *and a figure-eight knot* **(right)**

Let us begin with a question that is trivial but that can catch even mathematicians off guard. Tie an overhand knot in a piece of rope as is shown at the left in the illustration on this page. If you think of the ends of the rope as being joined, you have tied what knot theorists call a trefoil knot. It is the simplest of all knots in the sense that it can be diagrammed with a minimum of three crossings. (No knot can have fewer crossings except the trivial knot that has none.) Imagine that end *A* of the rope is passed through the loop from behind and the ends are pulled. Obviously the knot will dissolve. Now suppose the end is passed *twice* through the loop as is indicated by the broken line. Will the knot dissolve when the ends of the rope are pulled?

Most people guess that it will form another knot. Actually the knot dissolves as before. The end must go *three* times through the loop to produce another knot. If you try it, you will see that the new trefoil created in this way is not the same as the original. It is a mirror image. The trefoil is the simplest knot that cannot be changed to its mirror image by manipulating the rope.

The next-simplest knot, the only one with a minimum of four crossings, is the figure eight at the right in the illustration on this page. In this form it is easily changed to its mirror image. Just turn it over. A knot that can be manipulated to make its mirror image is called amphicheiral because like a rubber glove it can be made to display either handedness. After the figure eight the next-highest amphicheiral knot has six crossings, and it is the only 6-knot of that type. Amphicheiral knots become progressively scarcer as crossing numbers increase.

A second important way to divide knots into two classes is to distinguish between alternating and nonalternating knots. An alternating knot is one that can be diagrammed so that if you follow its curve in either direction, you alternately go over and under at the crossings. Alternating knots have many remarkable properties not possessed by nonalternating knots.

Still another important division is into prime and composite knots. A prime knot is one that cannot be manipulated to make two or more separated knots. For example, the square knot and the granny knot are not prime because each

DIGITAL



TOTAL CONTROL REMOTE CONTROL

INTRODUCING RCA's DIGITAL COMMAND CENTER. The master control for selected RCA ColorTrak 2000 TV's and compatible RCA video components. Its unique microcircuitry gives you total control over each separate component, letting you switch from broadcast to disc to VCR—instantly. Joined with a compatible RCA VideoDisc Player and VCR, like those shown at left, the Digital Command Center controls a total of 58 separate functions; VCR functions like single-frame advancement. And VideoDisc Player functions like pause and speed search. 58 functions, all at the touch of a button. See the RCA Digital Command Center for yourself at your RCA dealer. You'll see it's far and away the most impressive remote control you've ever laid hands on. For more information and a free copy of the "Living with Video" book ($2.50 retail value), write: RCA Consumer Electronics, Dept. 32-312N, P.O. Box 1976, Indianapolis, Indiana 46206.

WE'LL OPEN YOUR EYES. RCA



ACTUAL SIZE.

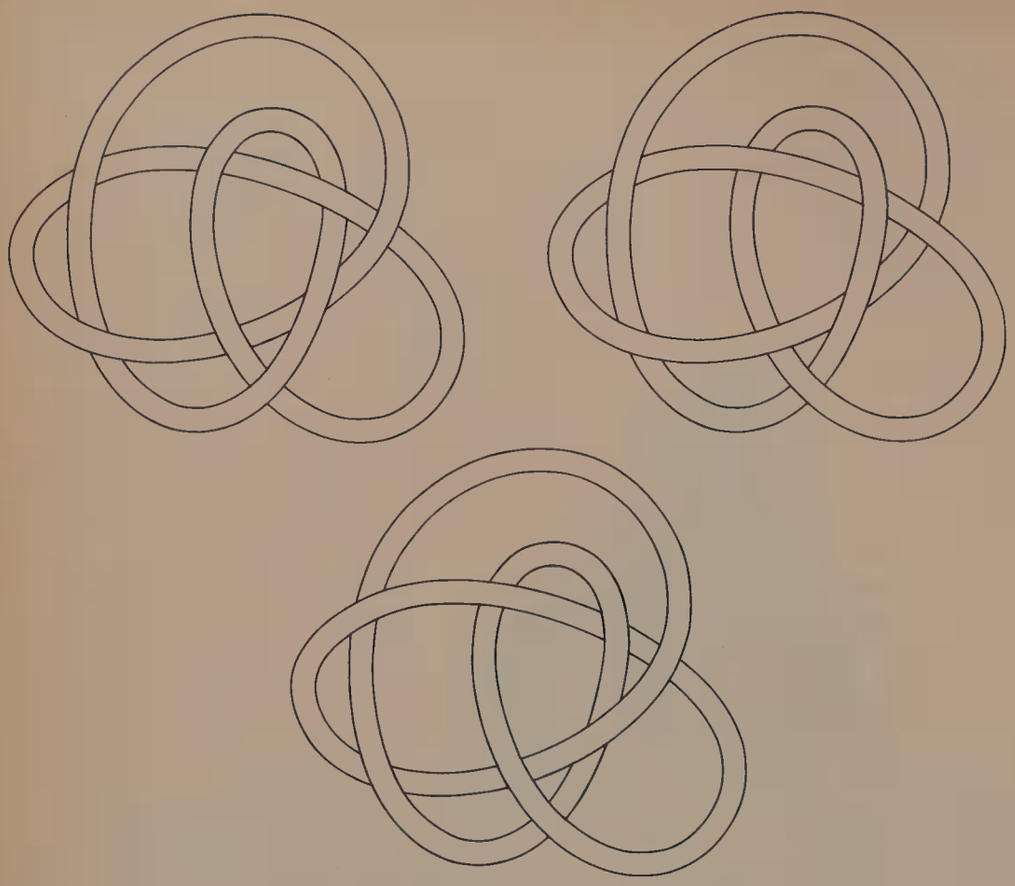
COMMAND


Only in

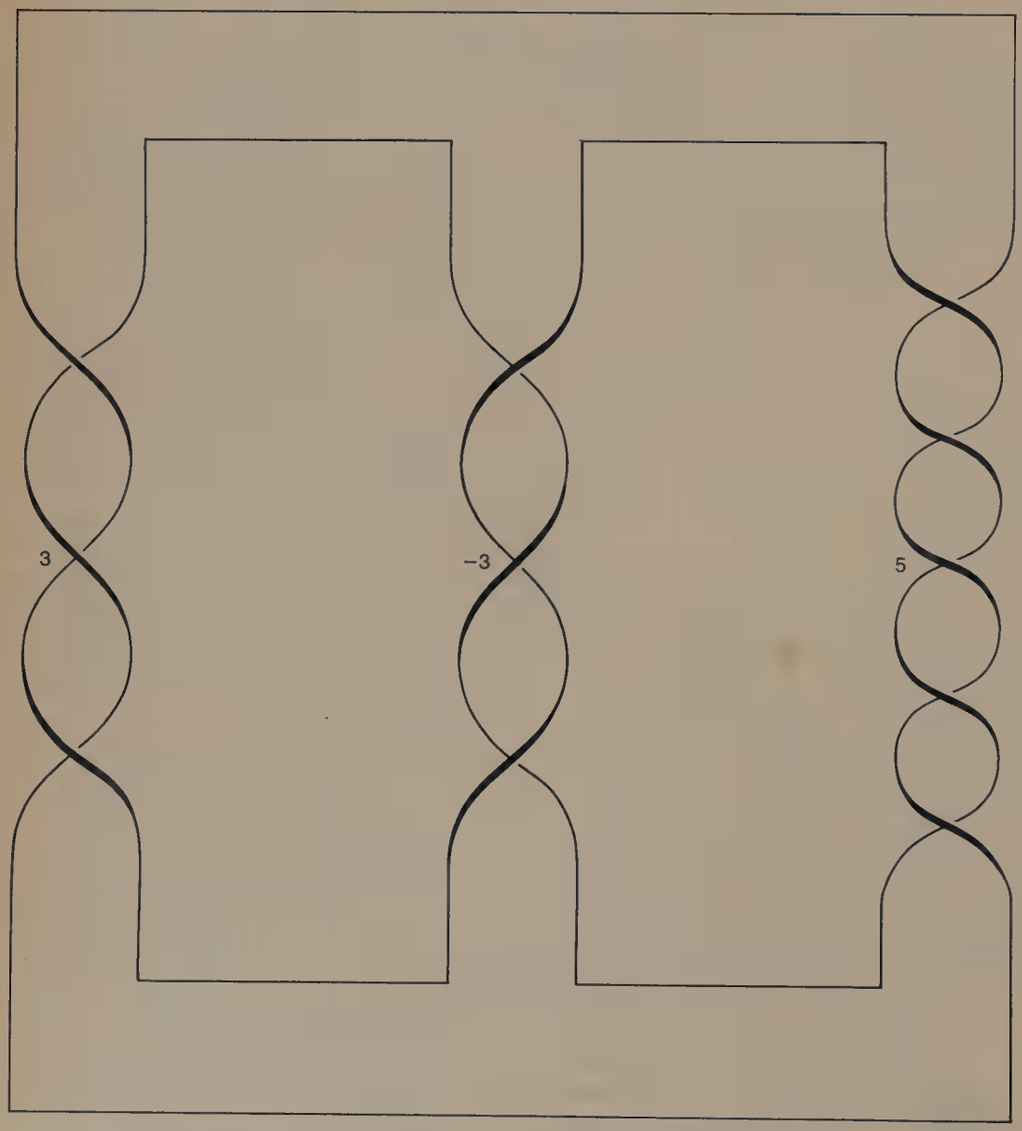


a Jeep.®

Jeep® CJ



The three nonalternating knots with eight crossings



The simplest member of an infinite family of noninvertible pretzel knots

can be changed to two side-by-side trefoils. The square knot is the "product" of two trefoils of opposite handedness. The granny is the product of two trefoils of the same handedness, and therefore (unlike the square knot) it is amphicheiral. Both knots are alternating. As an easy exercise to be answered next month, see if you can sketch a square knot with six (the minimum) alternating crossings.

All prime knots of seven or fewer crossings are alternating. Among the 8-knots only the three in the top illustration at the left are nonalternating. No matter how long you manipulate a rope model of one of these knots, you will never get it to lie flat in the form of an alternating diagram.

A fourth basic binary division of knots is into the invertible and noninvertible. Imagine an arrow painted on a knotted rope to give a direction to the curve. If it is possible to manipulate the rope so that the structure remains the same but the arrow points the other way, the knot is invertible. Until the mid-1960's one of the most vexing unsolved problems in knot theory was whether noninvertible knots exist. All knots of seven or fewer crossings had earlier been found invertible by manipulating rope models, and all but one 8-knot and four 9-knots. It was in 1963 that Hale F. Trotter, now at Princeton University, announced in the title of a surprising paper "Non-invertible Knots Exist" (*Topology*, Vol. 2, No. 4, pages 275–280; December, 1963.)

Trotter described an infinite family of pretzel knots that will not invert. A pretzel knot is one that can be drawn, without any crossings, on the surface of a pretzel (a two-hole torus). It can be drawn as is shown in the bottom illustration at the left as a two-strand braid that goes around two "holes," or it can be modeled by the edge of a sheet of paper with three twisted strips. If the braid surrounds just one hole, it is called a torus knot because it can be drawn without crossings on the surface of a doughnut.

Trotter found an elegant proof that all pretzel knots are noninvertible if the crossing numbers for the three twisted strips are distinct odd integers with absolute values greater than 1. Positive integers indicate braids that twist one way and negative integers indicate an opposite twist. Later Trotter's student Richard L. Parris showed in his unpublished Ph.D. thesis that the absolute values can be ignored provided the signed values are distinct, and that these conditions are necessary as well as sufficient for noninvertible pretzels. Thus the simplest noninvertible pretzel is the one shown. Its crossing numbers of 3, −3 and 5 make it an 11-knot.

It is now known that the simplest noninvertible knot is the amphicheiral 8-knot in the illustration on page 24. It was

How to make your *personal* computer more *personal*

Afraid? Of computers? Do they still seem cold, impersonal, and—well—just not right for you?

Don't feel bad. You're not alone. But at least there's something you can do to make those impersonal computers more personal. You can subscribe to Personal Computing magazine, America's guide to better computing.

Sound easy? It is, if you're serious about getting into computing. Here's how we can help: Personal Computing magazine can help you get started, keep you up-to-date, and help you stay current on the latest developments in hardware and software.

And, our articles are written for people who don't have advanced degrees in computer science. Personal Computing magazine is written in terms *you* can understand. In fact, it's like having yor very own computer consultant right in front of you. Only better. Because we're a magazine—easy to use, easy to digest at your own speed. Not a high-tech wizard who speaks another language. Best of all, you can keep every issue of Personal Computing to refer back to time and time again.

In Personal Computing, you'll learn what you can do with your computer—and just as important—what *else* you can do with it. We'll introduce you to people who have solved a wide range of problems with their computers. Problems you'll relate to, solved by *people just like you.* We'll tell you *what* they did, and *how* they did it. And we'll name names, quote prices, and show you where to go for more information. Getting into personal computing has never been easier. Or so painless.

Personal Computing is Answers. Every month, we answer the questions most frequently asked about personal computers. Questions about hardware, software, specific programs, what to pay, what to avoid, what certain systems can and cannot do, and more.



Personal Computing is Interviews. We talk to the people on the leading edge to find out what's on the horizon. The company presidents. The software writers. The wiz-kids who started it all. And the new generation who will take us even farther.

Personal Computing is Information. You'll find detailed listings of new hardware and software on the market in our monthly departments: *Hardware of the Month* and *Software of the Month,* the *only* complete source of information on everything that is newly available that month, with prices and addresses to write for more information—*you won't find this service anywhere else.* Plus, *Dealer Listings,* who has what you need in your area. And—exclusively for subscribers—*The Product Showcase,* a wide variety of hardware and software products that can be ordered directly from suppliers, to *save you money.* In addition, you'll discover a wealth of information in the advertisements you'll find Personal Computing, since every major computer company has chosen our magazine to reach you.

For business, home, or leisure use, computers are here—*now.* And more people—just like you—are getting involved every day. As a result, *Personal Computing is the fastest growing magazine in the world.*

Put Personal Computing to work for you. There's no better—or more inexpensive—way of leveraging your computer investment. Or making your personal computer more personal.

Send For A FREE ISSUE

We are so sure that Personal Computing is the very best way to get into personal computing and keep finding out what else you can do with your computer that we want you to try it, and see for yourself...on us.

To receive your FREE trial issue of Personal Computing, simply mail the attached postage-paid order card *today!* If order card is missing, please mail savings coupon below.

☐ ***Yes!*** *Rush me my FREE copy of Personal Computing and enter my subscription at the money-saving rate of only $14.97. That's 11 additional issues for a total of 12. I'll be saving over 58% on the single-copy cost and over 16% on the regular subscription rate!*

I understand that if I don't like the first issue. I can simply return your invoice marked "cancel" and I may keep the first issue, FREE.

Name *(Please print)*

Company Affiliation

Street

City/State/Zip

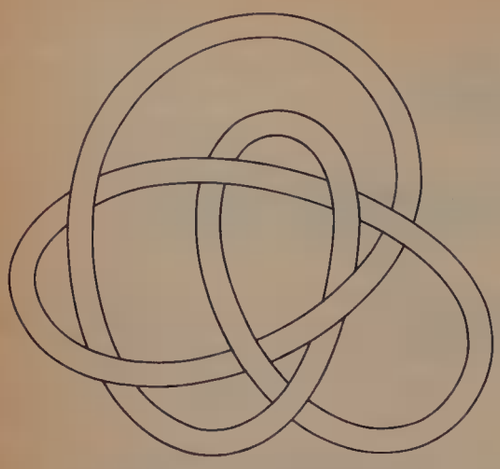
5AEJ6

PERSONAL COMPUTING

America's guide to better computing

P.O. Box 2941, Boulder, CO 80322

Please allow 4-6 weeks for delivery of your first issue. Your subscription may be tax deductible.



The simplest noninvertible knot

first proved noninvertible by Akio Kawauchi in *Proceedings of the Japan Academy* (Vol. 55, Series *A*, No. 10, pages 399–402; December, 1979.) According to Richard Hartley, in "Identifying Non-invertible Knots" (*Topology*, Vol. 22, No. 2, pages 137–145; 1983), this is the only noninvertible knot of eight crossings, and there are only two such knots of nine crossings and 33 of 10. All 36 of these knots had earlier been declared noninvertible by John Horton Conway, but only on the empirical grounds that he had not been able to invert them. The noninvertible knots among the more than 550 knots with 11 crossings have not yet been identified. In 1967 Conway published the first classification of all prime knots with 11 or fewer crossings. (A few minor errors were corrected in a later printing.) You will find clear diagrams for all prime knots through 10 crossings, and all links through nine crossings, in Dale Rolfsen's valuable 1976 book *Knots and Links*.

There are many strange ways to label the crossings of a knot, then derive an algebraic expression (or matrix) that will be the same for all diagrams of that knot. One of the earliest and most useful of such techniques gives what is called a knot's Alexander polynomial. Conway recently found a beautiful new way to compute a "Conway polynomial" that is equivalent to the Alexander one. If two knot diagrams give different algebraic expressions, they are different knots. Unfortunately the converse is not always true. Two knots may have the same polynomial yet not be the same. Finding a way to give any knot an expression that applies to all diagrams of that knot, and only that knot, is one of the major unsolved problems in knot theory.

Although there are tests for deciding whether any given knot is trivial, the methods are complex and tedious. For this reason many problems that are easy to state are not easy to resolve except by working empirically with rope models. For instance, is it possible to twist an elastic band around a cube so that each face of the cube has an under-over crossing as is shown at the left in the illustration on the opposite page. To put it another way, can you tie a cord around a cube in this manner so that if you slip the cord off the cube, the cord will be unknotted?

Note that on each face the crossing must take one of the four forms depicted in the illustration. This makes $4^6 = 4{,}096$ ways to wrap the cord. The wrapping can be diagrammed as a 12-knot, with six pairs of crossings, each pair of which can have one of four patterns. The problem was first posed by Horace W. Hinkle in *Journal of Recreational Mathematics* in 1978. In a later issue (Vol. 12, No. 1, pages 60–62; 1979–80) Karl Scherer showed how symmetry considerations reduce the number of essentially different wrappings to 128. Scherer tested each wrapping empirically and found that in every case the cord is knotted. This has yet to be confirmed by others, and no one has so far found a simpler way to attack the problem. The impossibility of getting the desired wrapping with an unknotted cord seems odd, because it is easy to twist a rubber band around a cube to put the under-over crossings on just two or four faces (all other faces being straight crossings), and seemingly impossible to do it on just one face, three faces or five faces. One would therefore expect six to be possible, but apparently it is not. It may also be impossible to get the pattern even if two, three or four rubber bands are used.

The illustration on page 26 depicts a delightful knot-and-link puzzle that was sent to me recently by its inventor, Majunath M. Hegde, a mathematics student in India. Note that the two trefoil knots form a granny. The task is to manipulate the rope and ring so that the ring is moved to the upper knot as is indicated by the broken line. All else must remain identical.

It is easy to do if you have the right insight, but it would spoil the fun if I did not postpone the solution until next month. Of course, the rope must not be untied from the chairs, nor are you allowed to open a knot and pass a chair through it. It will help if you think of the ends of the rope as being permanently fastened to a wall.

The trick of dissolving or creating knots by passing a person through a loop was actually used by fake mediums in the days when it was fashionable to relate psychic phenomena to the fourth dimension. Knots in closed curves are possible only in 3-space. In 4-space all knots dissolve. If you could toss an unknotted loop of rope to a creature in 4-space, it could tie any knot in the loop and toss it back to you with the knot permanently formed. There was a popular theory among physicists who believed in spiritualism that mediums had the power to move objects in and out of higher spaces. Some mediums, such as the American mountebank Henry Slade, exploited this theory by pretending to put knots into closed loops of cord. Johann C. F. Zöllner, an Austrian physicist, devoted an entire book to Slade and hyperspace. Its English translation, *Transcendental Physics* (Arno Press, 1976), is worth reading as striking testimony to the ease with which an intelligent physicist can be gulled by a clever conjurer.

Scientists are still being taken in by tricks involving knots and links. Psychic investigators William Cox and John Richards have recently been exhibiting a stop-action film that purports to show two leather rings becoming linked and unlinked inside a fish tank. "Later examination showed no evidence that the rings were severed in any way," wrote *National Enquirer* when it reported this miracle on October 27, 1981. I was reminded of an old conjuring stage joke. The performer announces that he has magically transported a rabbit from one opaque box to another. Then before opening either box he says that he will magically transport the rabbit back again.

It is easy, by the way, to fabricate two linked "rubber bands." Just draw them linked on the surface of a baby's hollow rubber teething ring and carefully cut them out. Two linked wood rings, each of a different wood, can be carved if you insert one ring into a notch cut into a tree, then wait many years until the tree grows around and through it. Because the trefoil is a torus knot, it too is easily cut from a teething ring.

The trick I am about to describe was too crude for Slade, but less clever mediums occasionally resorted to it. You will find it explained, along with other knot-tying swindles, in Chapter 2 of Hereward Carrington's *The Physical Phenomena of Spiritualism, Fraudulent and Genuine* (H. B. Turner & Co., Boston, 1907). One end of a very long piece of rope is tied to the wrist of one guest and the other end is tied to the wrist of another guest. After the seance, when the lights are turned on, several knots are in the rope. How do they get there?

The two guests stand side by side when the lights go out. In the dark the medium (or an accomplice) makes a few large coils of rope, then passes them carefully over the head and body of one of the guests. The coils lie flat on the floor until later, when the medium casually asks that guest to step a few feet to one side. This frees the coils from the person, allowing the medium to pull them into a sequence of tight knots at the center of the rope. Stepping to one side seems so irrelevant to the phenomenon that no one remembers it. Ask the guest himself a few weeks later whether

he changed his position, and he will vigorously and honestly deny it.

Roger Penrose, the British cosmologist, once showed me an unusual trick involving the mysterious appearance of a knot. Penrose invented it when he was in grade school. It is based on what in crocheting, sewing and embroidery is called a chain stitch. Begin the chain by tying a trefoil knot at one end of a long piece of heavy cord or thin rope and hold it with your left hand as in step 1 in the illustration on page 28. With your right thumb and finger take the cord at *A* and pull down a loop as in step 2. Reach through the loop, take the cord at *B* and pull down another loop (step 3). Again reach forward through the lowest loop, take the cord at *D* and pull down another loop (step 4). Continue in this way until you have formed as long a chain as possible.

With your right hand holding the lower end of the chain, pull the chain taut. Ask someone to select any link he likes and then pinch the link between his thumb and forefinger. Pull on both ends of the cord. All links dissolve, as expected, but when he separates his finger and thumb, there is a tight knot at precisely the spot he pinched!

A few years ago Joel Langer, a mathematician at Case Western Reserve University, made a remarkable discovery. He found a way of constructing what he calls "jump knots" out of stainless-steel wire. The wire is knotted and then its ends are bonded. When it is manipulated properly, it can be pressed flat to form a braided ring. Release pressure on the ring; tension in the wire causes it to spring suddenly into a symmetrical three-dimensional shape. It is now a frustrating puzzle to collapse the wire back to its ring form.

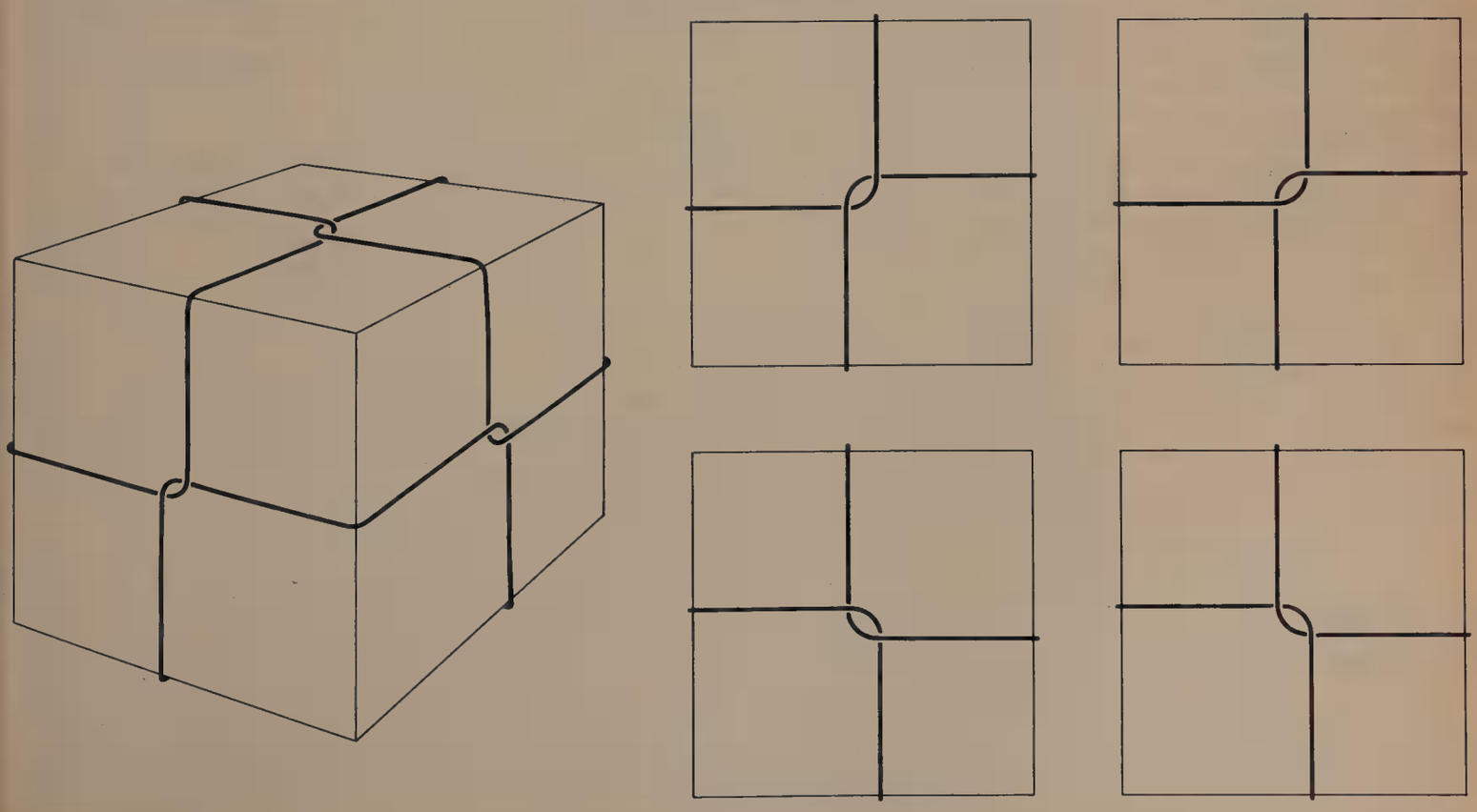
In 1981 Langer and his associate Sharon O'Neil formed a company they call Why Knots. From it you can obtain three handsome jump knots: the Figure Eight, the Chinese Button Knot and the Mathematician's Loop. When you slide one of these wire knots out of its square envelope, it pops into an elegant hanging ornament. The figure eight is the easiest to put back into its envelope. The Chinese button knot (so called because it is a form widely used in China for buttons on nightclothes) is more difficult. The mathematician's loop is the most difficult.

It was James W. Alexander (of the Alexander polynomial) who proved that every knot can be diagrammed as a closed circular braid. This property of all knots (not just torus knots) is known as Alexander's theorem. It is not easy to prove, nor is it easy, when you are given a knotted rope with many crossings, to find a way of altering it to a ring braid with the fewest strands.

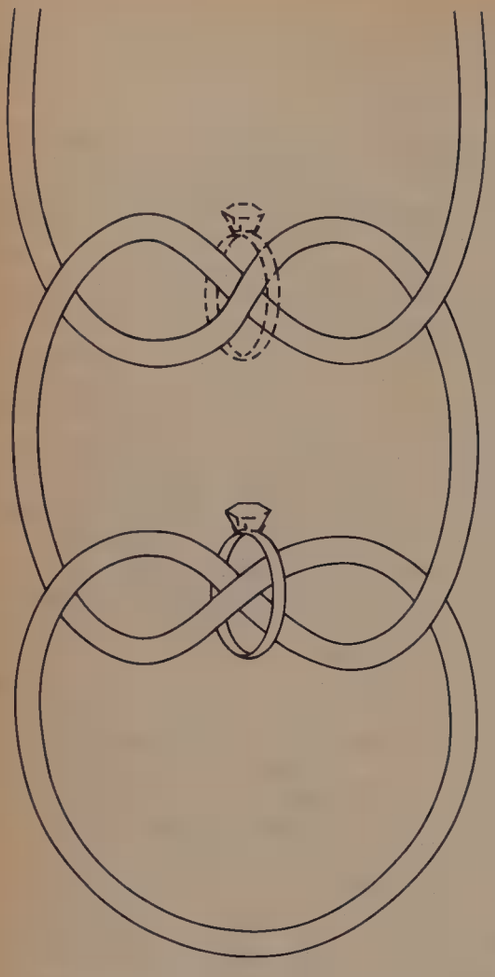
Langer tells me that anyone in the U.S. can get his three jump knots by sending $10.50 to Why Knots, P.O. Box 635, Aptos, Calif. 95003. These shapes make it easier to understand how the 18th-century physicists could have developed a theory, respectable in its day, that molecules are different kinds of knots into which vortex rings of ether (today read "space-time") get themselves tied. Indeed, it was just such speculation that led the Scottish physicist P. G. Tait to study topology and conduct the world's first systematic investigation of knot theory.

Douglas Hofstadter writes: The Luring Lottery, proposed in my June column, created quite a stir. Let me remind you that it was open to anyone; all you had to do was send a postcard with a clearly specified positive integer on it stating how many entries you wanted to make. This integer was to be, in effect, your "weight" in the final drawing, so that if you wrote "100," your name would be 100 times as likely to be drawn as that of someone who wrote "1." The only catch was that the cash value of the prize was inversely proportional to the sum of all the weights received by June 30. Specifically, the prize to be awarded was $1,000,000/*W*, where *W* is the sum of all the weights sent in.

The Luring Lottery was set up as an exercise in "cooperation" v. "defection." The basic question for each potential entrant was: "Should I restrain myself and submit a small number of entries, or should I 'go for it' and submit a large number? That is, should I cooperate, or should I defect?" Whereas in the examples of cooperation v. defection I had given earlier there was a clear-cut dividing line between cooperators and defectors, here it seems there is a continuum of possible answers, hence of "degree of cooperation." Clearly one can be an extreme cooperator and voluntarily submit nothing, thus in effect cutting off one's nose to spite one's face. Equally clearly one can be an extreme



The problem of the cube and the rubber band



A ring-and-granny puzzle from India

defector and submit a giant number of entries, hoping to swamp everyone else out but destroying the prize in so doing. Still, there is a lot of middle ground between these two extremes. What about someone who submits two entries, or one? What about someone who throws a six-sided die to decide whether or not to send in a single entry?

Before I go further it would be well to present my generalized sense of the terms "cooperation" and "defection." As a child, you often had adults admonish you for walking on the grass or for making noise, saying, "Tut tut tut—just think if *everyone* did that!" This is the quintessential argument used against the defector, and it serves to define the concept: A defection is an action such that if everyone did it, things would clearly be worse (for everyone) than if everyone refrained from doing it, and yet it is an action that tempts everyone, since if only one individual (or a small enough number) did it while others refrained, life would be sweeter for that individual (or select group).

Cooperation, of course, is the other side of the coin: the act of resisting temptation. It need not be the case, however, that cooperation is passive whereas defection is active. Often it is the exact opposite: the cooperative option may be to participate industriously in some activity, whereas defection is to lie back and accept the sweet things that result for everyone from the cooperators' hard work.

Typical examples of defection are blaring loud music through the entire neighborhood on a fine summer's day; not worrying about speeding through a four-way stop sign, figuring the people going in the crosswise directions will stop anyway; not being concerned about driving everywhere alone in your car, figuring there is no point in making a sacrifice when other people will just continue to guzzle gas anyway; not worrying about conserving water in a drought, figuring "everyone else will"; not voting in a crucial election, saying, "One vote can't make any difference"; not worrying about having 10 children in a period of rapid population growth, leaving it to other people to curb their reproduction; not devoting any time or energy to pressing global issues such as the arms race, famine, pollution, diminishing resources and so on, saying, "Oh, it's so depressing, but there's nothing one person can do."

When there are large numbers of people involved, individuals may not realize that their own seemingly idiosyncratic decisions are likely to be quite typical and to be re-created many times over, even on a grand scale; thus what each couple feel to be their own isolated and private decision (conscious or unconscious) about how many children to have can turn into a baby boom. Similarly, "individual" decisions about the futility of working actively toward the good of humanity amount to a giant trend of apathy, and this multiplied apathy translates at the group level into insanity. In other words, greed or apathy at the typical-individual level can translate into insanity or catastrophe at the mass level.

Garrett Hardin, the evolutionary biologist, wrote a famous article about this type of phenomenon, called "The Tragedy of the Commons" (*Science*, Vol. 162, No. 3859, pages 1243–1248; December 13, 1975). His view was that there are two types of rationality, one (call it the "local" type) striving for the good of the individual and the other (the "global" type) striving for the good of the group, and that these two types are in inevitable and eternal conflict. I would agree with his assessment, provided the individuals are unaware of their joint plight and are blindly carrying out their actions as if in isolation.

If, however, they are fully aware of their joint situation and yet in the face of it blithely continue to act as if their situation were not a communal one, then I would maintain they are acting with total irrationality. In other words, with an enlightened citizenry "local" rationality is not rational, period. It is damaging not only to the group but also to the individual. For example, people who defected in the One-Shot Prisoner's Dilemma situation I described in June did worse than if everyone had cooperated.

This was the central point of my June column, in which I wrote about "renormalized rationality," or "superrationality." Once you know you are a typical member of a class of individuals you must act as if your individual actions were to be multiplied manyfold, because they inevitably will be. In effect, to sample yourself is to sample the field, and if *you* fail to do what you hope the rest will do, you will be very disappointed by the rest as well. Hence it pays a lot to think carefully about one's situation in the world before defecting, that is, jumping to do the naively selfish act.

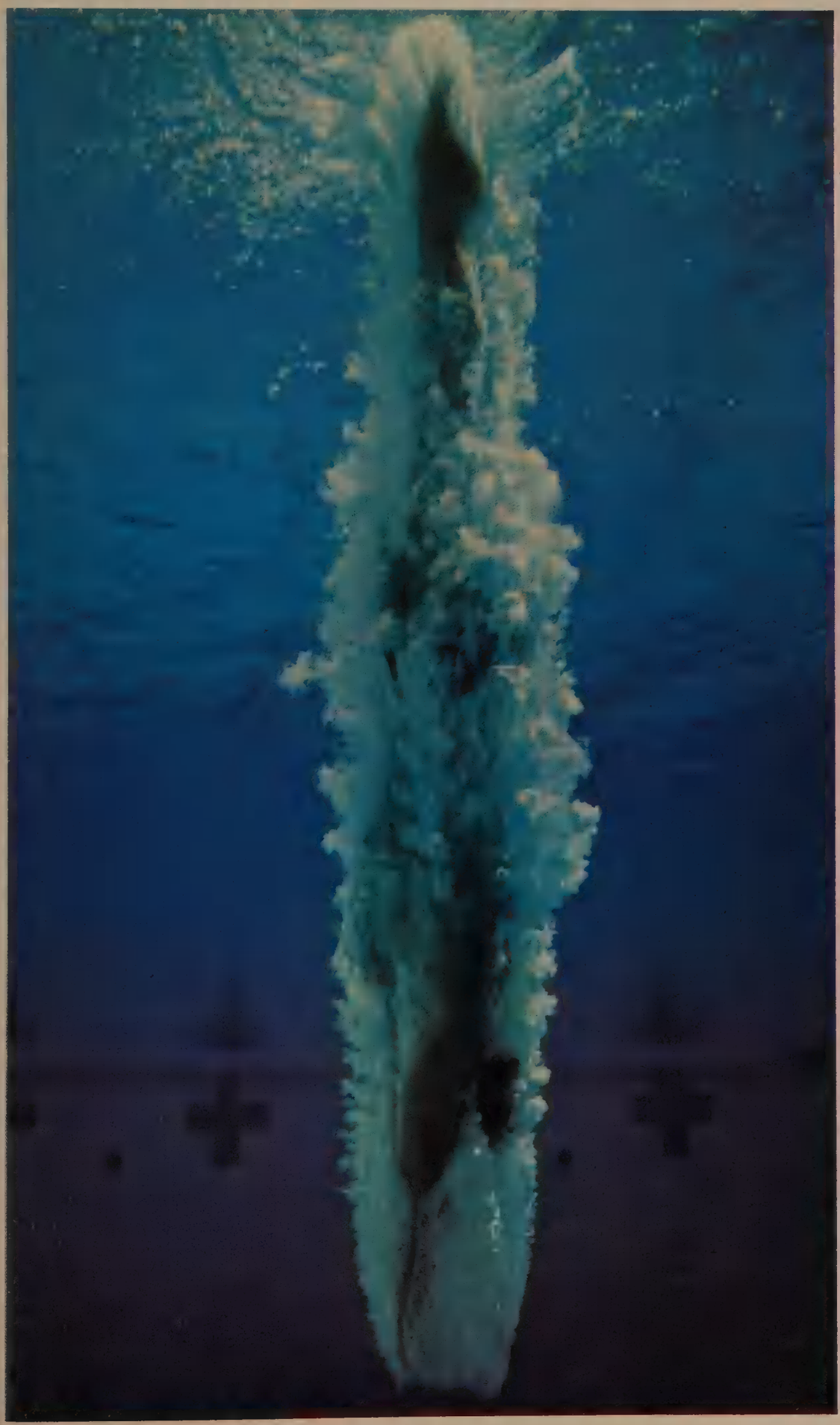
People strongly resist seeing themselves as a part of statistical phenomena, and understandably so, because it seems to undermine their sense of free will and individuality. Yet it is true that each of our "unique" thoughts is mirrored a million times over in the minds of strangers.

Nowhere was this better demonstrated than in the response to the Luring Lottery. It is hard to know precisely what in this case constitutes the "field." It was declared universally open, to readers and nonreaders alike. We would be safe, however, in assuming that few nonreaders would become aware of it, so let us start with the readership of *Scientific American*, which is at least a million. (The circulation of the English-language edition is 660,000.) Let us say that 100,000 readers read the column, and that 10,000 readers read it carefully and pondered the issues seriously. In any case I shall take this last figure as the population of the field.

In my June column I spelled out plainly the superrational argument that applies to the Platonia Dilemma for rolling an n-sided die and entering a lottery only if it came up on the right side. Here a similar argument applies.

In the Platonia Dilemma, where more than one entry is fatal to all, the ideal die turned out to have $1/N$ faces, where N is the number of players; hence 10,000 players represents a 10,000-sided die. In the Luring Lottery the consequences of more than one entry are not as drastic. Thus the ideal number of faces on the die turns out to be about two-thirds as many, and with 10,000 players a 6,667-sided die would do admirably. Giving the die fewer than 10,000 sides of course slightly increases each player's chance of sending in one entry. This is to make it quite likely that at least one entry will arrive!

With 6,667 faces on the die each superrational player's chance of winning is not quite one in 10,000; it is more like one in 13,000. The reason is that there is about a 22 percent chance no one's die will land right, so that no one will send in any entry and no one will win. But if you give the die still fewer faces, say 3,000, the expected size of the pot gets consid-



Everybody starts at the bottom.

Real winning comes in not staying there. In United States Diving, young people learn that it's okay to be afraid, that there's no shame in failing. And as they learn to dazzle the air and knife the water, they learn something much more important. That the most soaring triumphs are in simply trying again.

PHILLIPS 66

Phillips Petroleum has sponsored United States Swimming for eleven years, and United States Diving for five, helping with operating costs of national and international competitions leading to world championships. We continue to support them. Because their ideals touch the lives of everyone.

erably smaller, since the expected number of entrants gets larger. And if you give the die more faces, say 20,000, you run a considerable risk of having no entries at all. Accordingly there is a tradeoff whose ideal solution can be calculated without too much trouble, and 6,667 faces turns out to be about optimal, with the expected value of the pot being maximal: nearly $520,000, an amount not to be sneered at.

Now, this means that if everyone had followed my example in the June column, I would probably have received a total of one or two postcards with "1" written on them, and one of those lucky people would have won a huge sum of money. But do you think that is what happened? Of course not! Instead I was inundated by postcards and letters from all over the world: roughly 2,000 of them. What was the breakdown of entries? Part of it is given in this table:

1: 1,133
2: 31
3: 16
4: 8
5: 16
6: 0
7: 9
8: 1
9: 1
10: 49
100: 61
1,000: 46
1,000,000: 33
1,000,000,000: 11
602,300,000,000,000,000,000,000 (Avogadro's number): 1
Googol (10^{100}): 9
Googolplex (10^{googol}): 14

Curiously, many if not most of the people who submitted just one entry patted themselves on the back for being "cooperators." Stuff and nonsense! The *real* cooperators were those among the estimated 10,000 avid readers who calculated the proper number of faces of the die, used a random-number table or something equivalent, and—most likely—rolled themselves out. Perhaps among those several hundred single entries there was one that came from a lucky superrational cooperator, but I doubt it. The people who simply withdrew *without* throwing a die I would characterize as well-meaning but a bit lazy, not true cooperators—something like people who simply contribute money to a political cause but then do not want to be bothered any longer about it. It is the lazy way of claiming cooperation.

By the way, I have not by any means finished with my score chart. It is, however, a bit disheartening to try to describe what happened. Basically it is this. Scores of readers strained their hardest to come up with inconceivably large numbers. Some filled their entire postcard with tiny 9's, others filled their card with rows of exclamation points, creating iterated factorials of gigantic size. A handful of people carried this game much further, recognizing that the optimal solution avoids all pattern (for the reason see Gregory J. Chaitin's article "Randomness and Mathematical Proof," *Scientific American*, Vol. 232, No. 5, pages 47–52; May, 1975). It consists simply of a "dense pack" of definitions built on definitions, followed by a final line in which the "fanciest" of the definitions is applied to a relatively small number such as 2 or, better yet, 9.

I received, as I say, a few such entries. Some of them exploited such powerful concepts of mathematical logic and set theory that to evaluate which one was the largest became a serious problem, and it is not even clear that I, or for that matter anyone else, would be able to determine which is the largest integer submitted. I was strongly reminded of the lunacy and pointlessness of the current arms race, in which two sides vie with each other to produce arsenals so huge that not even teams of experts can say which one is the larger—and meanwhile the entire effort is to the detriment of *everyone.*

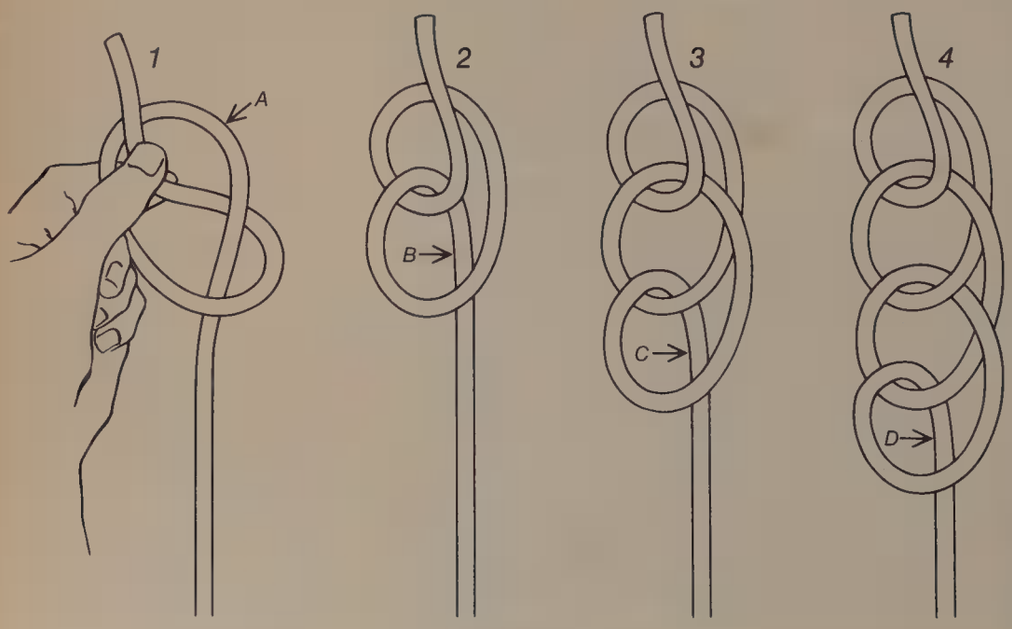
Did I find this amusing? Somewhat, of course. At the same time I found it disturbing and disappointing. Not that I had not expected it. Indeed, it was precisely what I had expected, and it was one reason I was so sure the Luring Lottery would be no risk for the magazine.

This shortsighted race for "first place" reveals how people in a huge crowd erroneously consider their own fancies to be unique. I suspect that most of those who submitted a number above 1,000,000 actually believed they were going to be the *only* one to do so. Many of those who sent in numbers such as a googolplex or 9, or 9 followed by thousands of factorial signs, explicitly indicated that they were sure they were going to "win." And then those people who pulled out all the stops and sent in definitions that would boggle most mathematicians were *very* sure they were going to win. As it turns out, I don't know who won, and it doesn't matter, since the prize is zero to such a good approximation that even God wouldn't know the difference.

Well, what conclusion do I draw from all this? Nothing too serious, but I do hope that it will give readers pause for thought the next time they face a "cooperate or defect" decision, which will probably happen within minutes for each of them, since we all face such decisions many times a day. Some of them are small, but some will have monumental repercussions. The future of the globe is in your hands.

With this perhaps sobering conclusion I should like to draw to a close my term as a columnist for *Scientific American*. It has been a valuable and beneficial opportunity for me. I have enjoyed the platform from which to express my ideas and concerns, I have (sometimes) enjoyed receiving the huge shipments of mail forwarded to me from New York several times a month and I have certainly been happy to make new friends through this channel. I shall not miss the monthly deadline, but I shall undoubtedly come across ideas that would have made good articles in "Metamagical Themas." I shall keep them in mind and maybe at some future time shall write a similar set of essays.

But for now it is time for me to move on to other territory: I look forward to a return to my professional work and to a more private life. Good-bye, and best wishes to you and to all other readers of this magazine, this issue, this copy, this piece, this page, this column, this paragraph, this sentence, this line and, last but not least, this "this."



Penrose chain stitch

ENERGY FUEL POWER

CHOICES FOR AMERICAN INDUSTRY AND GOVERNMENT



SCIENTIFIC AMERICAN

ADVANCED TECHNOLOGY CENTERS USA / SPECIAL ADVERTISING SUPPLEMENT/SEPTEMBER 1983



Rural America is Powerful Country

Across the sweep of rural America today, the countryside hums with activity.

It's a diverse and changing land. Supplier of food and raw materials for a nation . . . room for expansion of commerce and industry . . . new horizons for crowd-weary people.

It's our home. We grew up with its traditions and challenges.

We were busy turning on the lights across the countryside while most were calling it the "impossible dream." We knew even then we were taking on a commitment to the future—to provide adequate, dependable power to keep pace with changing needs.

Today, we own and maintain nearly half the nation's power lines to serve more than 25 million people in 46 states. In 1982 alone we hooked up more than 390,000 meters . . . energized nearly 27,000 new miles of line . . . made countless system improvements to serve our consumer-owners.

We know first-hand the challenges of today's rural America. We're at the center of activity in literally thousands of communities, providing reliable electric service, using our skill and know-how to keep things moving.

We have a big job . . . we work for a big and powerful country.

RE America's consumer-owned rural electric systems

For further information contact
NATIONAL RURAL ELECTRIC COOPERATIVE ASSOCIATION
1800 Massachusetts Ave NW Washington DC 20036
Telephone (202) 857-9540

It has been ten years. The long lines at the pumps are gone, as are "odd-even" license-plate days. The government no longer debates rationing, or even allocation.

The atmosphere of crisis—to which Americans responded, as we always do, with sacrifice and common sense—is past. But is the actual crisis past—or is it with us still, temporarily masked by a softened industrial economy and the lasting effects of massive American consumer conservation? And if the "energy crisis" is indeed as much or more a concern in 1983 as it was in 1973, what does the future hold for Americans, our industry, and our government?

NEED FOR USEFUL INFORMATION

The American businessman—and, broadly, Americans and their governments—must seek information to be used to guarantee energy/fuel/power security and sufficiency in future years, no matter what economic, social, or political disruption might occur. Energy for fuel, for heat, and for transportation are crucial to corporate economic survival. Essential as well are fossil-fuel product components—chemicals, resins, gases for products and their packages. The American "style" of management is effective and world-respected for making swift, rational decisions based on accurate, timely information. Yet, in regard to energy, that information can be stubbornly difficult to obtain.

Energy is a Babel of dissimilarity. Units of measure encompass barrels of oil, kilowatt-hours of electricity, cubic feet of natural gas, carloads or tons of coal. One simple, manageable way to regard energy data is to use the "quad" as a common denominator. "Quad" is short for the number quadrillion (that is, "1" followed by fifteen zeroes or 10^{15}; a million billions). A quad represents one quadrillion British thermal units (a Btu represents the quantity of heat required to raise the temperature of one pound of water by one degree Fahrenheit).

Reducing energy data to quads permits expressing America's sources and uses of energy, past, present, and future, in manageable terms. In 1980, 78 quads of primary energy were consumed in America to provide 60 quads of power for our residential, commercial, industrial, and transportation sectors (according to the U.S. Department of Energy). DOE projects that, in 2000, America will consume 100 quads to provide 68 quads of end-use power. For each year, the number of quads "lost"—18 in 1980, 32 in 2000—are expended in burning fossil fuels for generation of electricity, or in the transmission of electricity across power lines. Of the original 78 quads consumed in 1980, about 14 were imported. As a comparison, America consumed about 30 quads/year just after World War II, and was (for the last time) a net exporter of oil.

Predicting future energy consumption and needs has long since become more art than science. The U.S. DOE ventures to project, rather than predict. Mobil Oil Corporation, which does not publicize its future projections (which it terms "planning assumptions"), nevertheless between 1979 and 1982 revised its own estimates (of U.S. consumption in 2000) downward by 27% for oil consumption, and by 14% for energy consumption overall. Yet there remains agreement that consumption will, in 2000, be much higher than today, and of a different sort. If DOE is near the mark, then in 2000—when today's first-graders graduate from Harvard Business School or Wharton and begin to manage our economy—they will find that economy driven 65% by oil and gas, 18% by electricity, 11% by coal, and 6% by hydro and all other renewable power resources. The share of oil and gas in that economy will be 20% less than in 1980; electricity's share will be 50% greater, and the shares of coal and renewables will have doubled.

Where will 2000's energy—68 quads of it—be consumed? DOE projects 40% will be consumed in the residential/transportation sectors—down from the 1980 share of 50%. Commercial consumption will remain at its 1980 share. Industrial consumption, however, will rise from about 39% (of 1980 consumption) to 46% in the year 2000. Virtually all of the increased end-use quad consumption—68 quads overall in 2000, vs. 60 quads in 1980—will be consumed by industry. (It should be noted that electricity is projected to continue to consume more than three quads of fossil-fuel energy for each quad of power it ultimately generates.)

THE SCENARIO APPROACH

To further organize the information America's managers need for effective planning, a "scenario" approach to the major energy resources may help. The major resources are crude oil, natural gas, coal, electricity, and such renewables as solar and wind energy and hydropower. America may expect, in 2000 and beyond, a sufficient and secure supply of energy overall, and of most of the major resources. Yet a single American resource, not included above, is already in critically short supply.

That resource is time.

COAL—OUR PLENTIFUL RESOURCE

America holds one-fourth of the world's coal reserves—and the world's reserves are enormous. America's share aggregates about 200 billion tons or over 5,000 quads, according to The British Petroleum Company's "Statistical Review of World Energy 1982." At our 1980 rate of consumption, recoverable reserves will last more than two centuries. Total worldwide coal reserves, including those unlikely to be economically recoverable, could approach an astronomic 200,000 quads. The future seems assured, in terms of proved American reserves, and recent rates of consumption.

Still, what's removed cannot be restored. Coal is exhaustible, and more than three quads of coal are required to generate one quad of electric power. Should more and more coal be allocated to electricity, rates of coal consumption will rise. Should the cost of coal liquification and gasification become competitive with, or less than, other fossil fuels, consumption will rise.

America's managers in the next 15 to 20 years will find coal an available resource for heavy manufacturing and an available base resource for generation of electricity.

Coal is mined far from the furnaces it feeds. Our rail transportation system has been massively upgraded in recent years, and offers excess capacity at present. The railroad industry today has facilities to deliver double its current coal freight business, and can adapt to almost any projected increase in coal loadings through the century's end. It cannot, however, efficiently deliver ever-larger quantities without increased capital costs and, eventually, necessarily higher freight rates. Furthermore, coal is an environmentally costly energy resource. Burned at utilities in huge amounts, it will give off problematic gaseous and particulate matter even if combustion efficiency is very high. And the cost of developing scrubbers and similar procedures to increase that efficiency, or to reduce unwanted post-combustion material, will be reflected in electric rates.

POWER IS THE KEY TO AMERICA'S FUTURE

By
David Packard
Chairman of the Board, Hewlett-Packard Company

It is time, I think, for this country to adopt a sane, forward-looking approach to meeting our energy needs. We must replace emotion with reason and abandon quick-fix responses for long-term solutions.

We must define our long-term energy goals and set about making the commitments and investments that are essential to achieving those goals. No area is in need of greater attention than the nation's electricity supply.

Basic and New Industry

Electricity is a key factor in any long-range energy policy. Our basic industries are becoming increasingly electrified. Our new, high technology industries are all electronic based. Not only will our use of electricity increase substantially with a step-up in industrial activity, increased electricity use and supply are essential to the nation's sustained economic recovery.

"If we continue in the current trend, not only will sufficiency of supply be threatened, but cost to consumers will be higher than need be."

The link between growth in electricity use and growth in Gross National Product has been established over many years. This linkage has remained even as we have become more energy conscious and have reduced the overall energy content of manufactured products.

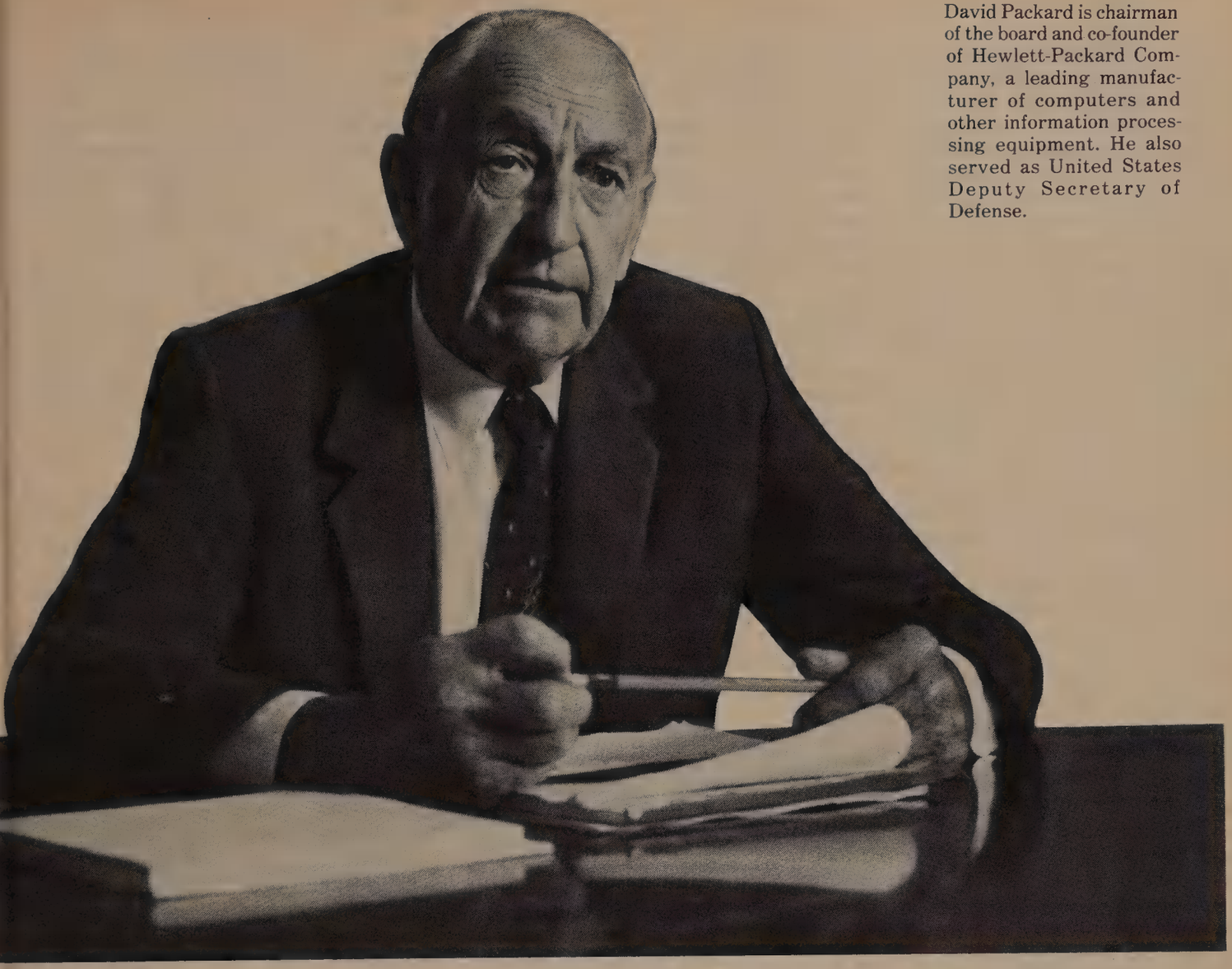
In many industries, electricity is replacing other energy forms that are less efficient at the point of use or are subject to the whims of foreign suppliers. This trend undoubtedly will—indeed, must—continue.

At the same time, our means of generating electricity must be made more efficient and secure. In recent years, electric utilities have made great strides in displacing imported oil as a fuel for generating power. But much remains to be accomplished. Electric utilities are still the largest stationary users of oil in the nation.

Danger of Complacency

We must resist being lulled into false complacency by declines in energy use and the resulting appearance of excess supplies that have accompanied economic recession. Any such cushion of supply will be quickly eliminated by restoration of economic vigor.

Right now, new power plants must be planned and commitments made for construction up to 14 years before they can be put into service. Decisions made today will determine whether we will be able to power American industry reliably and economically in the decades ahead. Today's decisions will determine job opportunities for the generation now growing to adulthood, our ability to compete for world markets with



David Packard is chairman of the board and co-founder of Hewlett-Packard Company, a leading manufacturer of computers and other information processing equipment. He also served as United States Deputy Secretary of Defense.

ations that are still developing eco-
omically, and the basic security of
ne country as we head toward the
1st century.

Unfortunately, most of the decisions
eing made today are weighted
oward not building for the future.

New Construction Risky

Utility managements, responsible
or preserving the savings of those
who have invested in their companies,
annot risk building for tomorrow
when they are unable to earn com-
petitive returns on plants in place
oday. They are being forced to oper-
te on the basis of short-term sur-
ival after a century of being America's
ost long-term oriented industry.

Solutions to the utility indus-
ry's financial problems and the
ation's long-term energy problems
coincide:

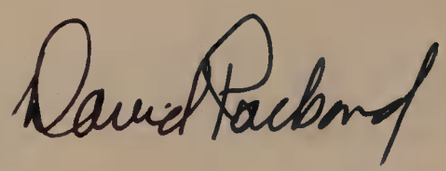
. Rates charged for electric power must reflect the actual cost of providing service and provide a reasonable return to shareholders.
2. The licensing and regulatory process must be streamlined so that new generating plants can be brought into service more quickly.

Combined, these steps will assure the nation sufficient power at lower cost in the future. If we continue in the current trend, not only will sufficiency of supply be threatened, but costs to consumers will be higher than need be.

Everyone Responsible

Who is responsible for developing those solutions? While regulators at the federal and state level make the final decisions that determine rates of return and power plant licensing procedures, we cannot expect that they alone can correct problems with the regulatory process.

Government, in the long run, reflects the attitudes of the people. It is up to all of us—business, industry, politicians and consumers—to demonstrate our support for policies that will enable us as a nation to exercise sound options for the future.



David Packard

This message is sponsored by the Edison Electric Institute, which represents the investor-owned utilities that deliver 77% of the nation's electricity.

Participating independent authorities are not paid for these messages. They believe a healthy electric industry is of critical national importance. EEI welcomes your comments.

Edison Electric Institute
The association of electric companies
1111 19th Street, N.W., Room 716, Washington, D.C. 20036

In energy-resource planning, each alternative scenario invariably unveils new social and economic costs. Research is ongoing to develop more environmentally acceptable ways to mine, move, and burn coal. And the nation's largest synthetic fuel plant—a coal gasification facility—is set for completion in 1984, following resolution of some major short-term financial uncertainties related to Federal commitment.

Unless coal is liquified or gasified, it offers no practical portability or storability for use in motor vehicles or residential heating, or, in nonurban areas, in commercial establishments. So the American economy may, in 2000, be practically unable to quickly substitute coal, our most plentiful energy resource, for oil or natural gas. In 1925, coal supplied 70% of U.S. energy needs and about half our electricity. No similar role is projected for coal for generations to come.

OIL—ESSENTIAL TO TRANSPORTATION

In 1960, President Eisenhower enacted quotas to restrict imports of oil, and Americans were asked to drive an extra four minutes per day to reduce the national surplus. In 1964, one Esso gasoline retailer in East Providence, Rhode Island would "put a Tiger in your tank" at 19.9 cents per gallon. The automobile had become a symbol of American independence, and the economies of several foreign nations were becoming based on the feeding of just that symbol. Responding to the Eisenhower quotas, Venezuela initiated an association of these nations, to be termed OPEC; fifteen years later that acronym would become a fact of American economic life.

Worldwide economic oil reserves aggregate some 3,700 quads, perhaps 35 years' supply at 1981 rates of consumption. Recoverable American reserves currently amount to 150-200 quads, enough to meet domestic needs for between eight and 20 years (depending on our rate of consumption). Of course oil imports extend the number of years' supply available to us; past imports have allowed us to preserve the few domestic reserves we now have.

Exploration, conservation, and diversion are the most productive means of extending our crude-oil reserves. Importation is a short-range solution but our experience has been that importation can leave America vulnerable to substantial disruption.

Conservation, at least at the consumer level, has reduced oil consumption dramatically. No industry expert could have predicted the extent to which Americans would be willing to reduce both the size and speed of their family cars, to spend (with tax assistance) for energy-conserving home improvements, or to adjust home thermostats beyond preferred comfort levels. Yet consumer conservation, in the words of one venerable Yankee, "is a good dog that hunts just once." In the event of a future major oil-supply disruption, a similar and additional reduction of consumer oil use would reduce the average American standard of living to an unacceptable level. Commercial and industrial conservation, while offering potential oil savings, cannot be accomplished without time-consuming and expensive alterations to existing plants and equipment.

Diversion—the substitution of a more plentiful energy resource for a scarce one—is a productive way to extend reserves. Using electricity for heating could free from 3 to 4 quads of end-use oil or gas for alternative use elsewhere in the economy. Using an alternative fuel to generate electricity could save an additional 3 quads—combined, an amount equal to half America's total 1980 oil imports.

Exploration, and the bringing-to-market of found oil, remains America's best prospect for oil sufficiency beyond 2000. The expense of exploration, expressed in terms of time and money, is enormous. America, and other nations, have characteristically consumed their least-expensive, most-readily-available energy resources. Conservation and diversion extend the life of American reserves, but only exploration can increase their size. In an address to world energy experts in Denver last November, Tony Scanlan, Economic Adviser to BP Oil said, "I hope (your new offshore West Coast oil discoveries) will continue, but you will need the equivalent of a new East Texas field every year just to maintain reserves. Twice in the past decade the world economy paid a very high, an unnecessarily high price for failing to heed the simple warning that it must stop trying to live off its seed corn, marginal cost oil, and I fear we may be about to do so again . . . ."

Wellheads and transport routes on our side of the globe seem secure, and non-U.S. reserves seem promising. According to the *Oil & Gas Journal,* Mexico's proved reserves exceed the total of those of the U.K., Indonesia, and the rest of Europe and the Asia/Pacific regions combined, not including reserves added after January 1981. It would seem very sensible for America's managers to look realistically at future relationships with our southern continental neighbor.

The rest of our hemisphere contains three quads of proved oil reserves for each one of America's, and one quad of natural gas for each of our own. Should we, in a worst-case scenario, fail to plan adequately for domestic energy security and sufficiency, we will be well-served to take note of possible partnership with our near neighbors. We will be best-served to begin developing such partnerships, with fairness and respect on both sides, under any possible scenario.

NATURAL GAS—CLEAN, PORTABLE, EFFICIENT

Natural gas represents more than half of America's overall energy consumption in the residential and commercial sectors, and 40% of industrial consumption. For years the nation has been drawing down its proved reserves faster than it has been replacing them. The trend, however, is toward near equilibrium in production and consumption, at least in the Western Hemisphere; furthermore, consumption in 1981 vs. 1973 declined about one-half of one percent, as did production over the same eight-year period.

Deregulation has stimulated exploration, and all estimates indicate large reserves likely to be discovered within American territory. Residential gas prices in some Northeastern areas have inceased 20% or more since late 1982, an event which follows several years during which gas for heating was much less expensive than heating-oil, and during which many residences converted from oil to gas heat. If demand remains high, significant recent discoveries in the Rockies and Louisiana's Gulf coast will be brought in. And the Potential Gas Committee of the Colorado School of Mines has estimated that Alaska may hold nearly five times the reserves of natural gas proved to date.

Midrange projections, therefore, by the U.S. Department of Energy show a 25% increase in gas consumption by 2000, and should exploration and pipeline construction proceed in response to higher prices already allowed, the supplies will be there. Even in a "worst-case" scenario, America will presumably have access (at cost) to Mexico's huge reserves, and many energy-industry companies have devised workable (if expensive) technologies to synthesize natural gas from coal and other resources.

Like coal, natural gas offers the promise of near-term sufficiency and security of supply, although at ever-higher cost. Like coal, natural gas is ultimately exhaustible. Unlike coal, natural gas requires a much less expensive transportation and environmental-control infrastructure for delivery and combustion. Finally, natural gas shares with coal a basic lack of portability: it cannot, in its original state, meet America's need for a fuel which can be burned in transit.

ELECTRICITY—AMPLE, SECURE, CONTROVERSIAL

Electric utilities—regulated at the State level for sales to businesses and consumers, and at the Federal level for transactions among themselves—are consumers, converters, and distributors of energy. Where other sectors of the energy industry have seen increasing deregulation and more free-market flexibility, electric utilities are experiencing much greater amounts of regulation—at a rate the Department of Energy itself describes as "alarming."

DOE projects that electricity—fueled by a variety of base resources—will increase from about 12% of overall American power generation in 1980 to about 18% in 2000. Including losses of energy due to conversion and transmission, electricity will increase its consumption of all of our energy from about 25 quads in 1980 (32% of all energy consumed in America) to 44 quads in 2000 (44% of all energy consumed). In general, electricity is an enormous vortex consuming three to four quads of fossil fuels for each quad of power generated. In 1981, an amount equivalent to a third of total U.S. oil imports went to fire utility boilers. Two-thirds of this quantity is considered uneconomic at present oil prices. Separately, eighty percent of coal consumption is accounted for by electric utilities. Over the past ten years, coal consumption by electric generating plants has increased 82%.

At the beginning of the century, electric companies provided the nation's electricity needs at the least possible cost, receiving, in essence, natural monopoly rights to provide service within certain geographic areas. Each succeeding year electric costs fell as rising demand and improved technology contributed to industry development and effectiveness. This trend halted abruptly ten years ago.

Fossil-fuel costs skyrocketed at a time when rate-adjustment regulatory processes were already lengthy and politically sensitive. New environmental regulations added to construction and refitting costs. Highest of all was the cost of money: 30% of total annual manufacturing investment in America is consumed by electric utilities—and inflation, coupled with high interest rates, left the utilities vulnerable.

As a result, the utilities slowed their rates of investment in new plant and equipment, particularly when that plant and equipment might have been fueled by coal or oil or natural gas. Most new construction had been projected for plants fueled by atomic, rather than fossil-fuel, energy, and by 1973 there was a wide sense of relief that nuclear projects had been initiated as early (mid-Sixties) as they had.

In 1973, most sectors of the American economy were in favor of continued development of nuclear-fueled electricity. Industry recognized the value of a fuel source which could be adapted to many of its long-range needs and which was not subject to overseas politics. State and Federal governments were supportive. And fewer than three Americans in ten believed their communities would be more dangerous if a great increase in the number of nuclear power plants were to occur.

It goes without saying that things have changed. The General Electric Company has conducted energy-oriented market research consistently since 1964. Given that regulation drives electric power into the political arena, GE's summary of popular attitudes toward energy, based on the Company's proprietary research, is instructive:

> "The dire forecasts of the Club of Rome were first published (*Limits to Growth*) in 1972 when the general public was indeed pessimistic about the earth's supplies of natural resources. Since that time about half of that pessimism has disappeared, with twice as many respondents being skeptical about exhaustion of coal supplies and somewhat similar proportions becoming skeptical about exhaustion of oil and gas. Meanwhile, more respondents have learned (due to Three Mile Island) that, along with coal or oil, electricity can be made from nuclear energy.
>
> "While shortages are no longer considered likely, the abundance of resources, the increasing costs and dangers of power production, and a belief in the effectiveness of conservation have all influenced the rising view that new power plants are not needed. In short, skepticism about supplies has been replaced by skepticism about needs."

At present, American industry and government cannot ignore negative public opinion regarding nuclear power, and must plan for a future composed of "brownouts," insecurity, and high costs. The Department of Energy has stated flatly that "Nuclear power has proved to be a safe, economical, and environmentally acceptable energy source." Chairman Ian Ross of Bell Laboratories said in May before a Phoenix audience, "I don't think this nation can prosper as it should without well-engineered nuclear power plants." And yet 78% of Americans (based on a study published by *Public Opinion*) consider themselves at least moderately well-informed about nuclear power (a proportion larger, incidentally, than that of Americans who feel well-informed about their local school systems). So the reality of a "scenario" involving nuclear power is dependent on what the average American perceives that reality to be. John G. Kemeny, the Dartmouth educator who headed President Carter's commission investigating the Three Mile Island incident, concluded that "the average American ... hasn't the foggiest notion of most of our findings."

The Federal Government has direct responsibility for proper disposal of highly radioactive or very long-lived nuclear wastes, and for funding of research to develop "breeder-reactor" technology which would reprocess waste into reusable fuel. DOE states "there is no question that technically acceptable means of disposal exist," yet Americans in general do not believe this to be true. And while breeder research and development continues, along with research into nuclear fusion which would produce no waste at all, no prospect for practical commercial application is seen before the middle of the next century. In July 1981 DOE projected that, of 100 quads of energy production needed in the year 2000, more than 10 would be specifically nuclear-sourced, up from less than 3 quads in 1980. The two years since DOE's projections have seen nuclear reactor order cancellations rise to a hundred—and no new orders have been placed since 1978.

America's rural population is growing as fast as the urban population for the first time. As reported in SCIENTIFIC AMERICAN in July, "rural areas' attractiveness for industry is responsible sponsible for much of their population growth.... The new rural demographic concentrations appear to represent small centers of urban culture transplanted to the countryside and enabled to survive by recent advances in communications, transportation, and methods of industrial production." Such "transplants" have required shifts in energy resource planning, especially among electric utilities which

must project increases in consumption greater than those expected for the nation as a whole.

Fortunately much of that essential planning was undertaken several years ago, in large measure through the offices of the Federal Rural Electrification Administration. The REA has facilitated capital borrowing by rural electricity-generating and -distributing cooperatives, thus avoiding cutbacks in service to the new industrial population, and helping to manage local growth. It is an example of regional/Federal cooperation which can continue to benefit millions of nonurban Americans.

What is the reality, then, for business and government? In March of this year, the 64 electric companies comprising the New England Power Pool signed a supply agreement with Canada at a cost, over its term, of some $5 billion. The agreement is expected to save New England nearly 60 million barrels of oil which would have been required to fuel the Pool's boilers. Just as America turned to imports in the face of insufficient domestic production of oil after World War II, America has begun to consider imports of electricity as well.

RENEWABLES— THE LONG-RANGE HOPE

The sun is the generator of most renewable energy resources: the sun contributes to wind currents which drive windmills, gives light for plants which can produce methane, stovewood, and similar "biomass" energy, creates rain for use as hydropower, drives ocean currents for tidal/wave energy, and provides fuel for rooftop panels and similar heat-making processes.

It is difficult to measure renewables in terms of "quads" since, with the exception of hydropower, most renewables consume relatively large amounts of investment capital and provide energy only on a local level. Renewables contributed fewer than two quads of end-use power (by DOE estimate) in 1980, and are projected to provide about four of 68 quads of end-use power in 2000. Not included here is the hydropower primary energy which fueled about 11% of American electricity in 1981.

The individual industrial manager may divert a fraction of his energy/fuel requirements away from scarce or costly fossil fuels by means of installation of equipment permitting consumption of renewable energy. It must be realized, however, that such resources are usually intermittent, and thus not dependable for energy at the time, and in the quantity, that energy is wanted. In many parts of America, installation of equipment fueled by renewables will be in addition to, rather than in place of, eqiipment fueled by other resources.

The value of a discussion of renewables, as a potential solution to large-scale American energy needs, is in the reducing of wishfulness to practical information. Renewables are popularly perceived to be neither scarce nor insecure nor environmentally objectionable, and for those reasons many Americans hope for an expansion of the contribution they might make. The National Audubon Society has recently issued a sobering report on the environmental impact of, for example, intensive local construction of windmills. Noise levels would be high, and indigenous and migrating wildlife habits could be seriously disturbed.

In a *Newsweek* article in 1979, D.C. Winston of the Atomic Industrial Forum calculated that our 68-quad energy requirements in 2000 "could be met by covering an area equivalent to that of the state of Oregon with solar collectors." These collectors, he added, "will contain large—truly large—tonnages of cadmium, silicon, gallium, copper, arsenic, sulfur . . . of varying availability on world markets." And Winston suggests the resulting irony: America could face the prospect of consuming more energy to manufacture solar cells—or other equipment—than the products themselves will generate.

On a national basis, renewables as a basic resource involve many of the same uncertainties, scarcities, and high costs, as other resources. In a 1982 address, Texaco Chairman John McKinley stated his strongest recommendation for meeting America's expected energy needs: evaluate and develop all energy resources in a timely sequence while continuing basic research and development for the mixed-fuel economy of the future—using money and manpower in the most effective ways. Renewables alone cannot, and are not projected to, provide a cure-all for such energy/fuel/power needs as America will incur.

Cover Photo: Computer-graphics image of the sun's corona, based on a photograph made by a research group from the High Altitude Observatory at the National Center for Atmospheric Research, sponsored by the National Science Foundation with a camera developed by Gordon A. Newkirk, Jr.

"ENERGY/FUEL/POWER" was produced by Ruder, Finn & Rotman, Inc., Washington, D.C.

TIME TO SUCCEED

American industry and government may be assured of secure, sufficient resources of energy/fuel/power if Americans will invest the necesssary time to develop those resources. Coal is plentiful for many present needs and our transport infrastructure can service a marked increase in its use. Yet coal is exhaustible, and its cost will rise significantly if the coal industry is obliged to extract from less accessible reserves. And coal is transmutable into oil and natural gas at costs which, in the future, may become more attractive than they are at present.

On a continental and hemispheric basis, crude oil and natural gas reserves—both proved and expected—should last several decades. These fossil fuels will become ever scarcer in the near term, as exploration and development are postponed, and America is far from safe from a "replay" of the supply crisis of ten years ago.

The energy "scenario" which holds the most promise for all sectors of the American economy is that involving a safe, economic, and technologically-advancing atomic-powered electric-energy capability. All other scenarios involve risks to the American economy and standard of living, risks which are different from, but equivalent to, the perceived and actual risks of nuclear power. Substitution of nuclear-powered electricity for such fossil-fuel-intensive end-uses as home heating and mass transportation would permit diversion of coal, oil, and natural gas to uses for which they are uniquely applicable, while extending the span of years for which reserves of them would be proved. Substitution of nuclear power for fossil fuels in the generation of electricity would immediately treble the quad-reserve of each fuel, as it is projected to be consumed.

It is safe to say that Americans will not tolerate any reduction in electric power below the amount we require, and safe to say as well that America will consume vastly greater quantities of nuclear-generated electricity than we do now. Whether we generate that power ourselves, or import it at greatly higher economic cost, is a decision still ours to make. The production and consumption of energy is far and away the largest industry in America, and the primary determinant of our current and future standard of living. How we manage today's resources, and whether we plan for tomorrow's, will determine our future.

BOOKS

The Mediterranean desert, Christopher Wren, monstrous growths, and industrial chemistry

by Philip Morrison

THE MEDITERRANEAN WAS A DESERT: A VOYAGE OF THE *GLOMAR CHALLENGER*, by Kenneth J. Hsü. Princeton University Press ($17.95). The evidence that glaciers once covered much of central Europe, not to mention North America, is overwhelming. The idea is now a commonplace; every schoolchild learns it, if only from the comic strips. About 10 million cubic kilometers of the mineral ice were added to cap Europe alone over the past million years. The first insight into that preposterous past came from the Swiss peasants who lived where a thoughtful person could see almost in one glance both the moraines and boulders at the living glacier edge and their counterparts mysteriously strewn down along the pleasant valley where no glacier had ever been known to intrude. The geologists who incredulously verified that picture and extended it step by step a century and a half ago needed no equipment beyond stout shoes and perhaps an alpenstock and a hammer.

This personal account, its first draft written on shipboard, tells a parallel story but a modern one. A similar if somewhat smaller volume of the mineral water was not added but removed from the map of Europe about five million years back. The central wine-dark sea, where Odysseus made his finally happy voyage, was then a desert, a "deep, dry, hot hellhole," its wide basin three kilometers below sea level, the seat of whirlwinds stirring red painted deserts, of transient briny lakes and of long river-cut canyons. Ken Hsü first found clear evidence for this picture (by now of course much elaborated) on the morning of August 24, 1970, watching his partner, Bill Ryan, another young shipboard sedimentologist, working up the sand-and-gravel sample they had collected during the long sleepless night before. The gravel held gypsum, an evaporative residue of seawater, along with two or three other components typical of the gravelly outwash from a long-eroded bed formed in a dried-up salt lake. A keen eye, a wash bucket and a hot plate were all the laboratory equipment needed to disclose the ancient desert that is the only plausible source.

No one, however, had ever seen or walked out those beds of gravel that furnished the illuminating little sample. It was fished out of two kilometers of sea off the Barcelona coast, held in a core barrel that had drilled deep into the ooze of the sea floor until it had struck a harder surface below, long recognized solely by its ubiquitous echo reflection up to the sounding ships, now disclosed as desert floor. The bright flash of inference was in the lofty tradition of field geology back to Louis Agassiz and James Hutton; the tools were no longer the traditional ones but the very epitome of big science, heavy investment and teamwork on a grand scale.

A hammer can loosen a sample from a rock wall by hand. A collector underseas needs much more, an entire ship dedicated to deep-sea drilling. Such a ship was the *Glomar Challenger* (now ignominiously idle, awaiting a buyer), sponsored by the National Science Foundation, sailing on the 13th drilling cruise out of the 80-odd legs of its career of deep-sea drilling. The product of Texas offshore-oil technology, with an assist from the shadowy world of technical intelligence, it was the concrete embodiment of that half-dream and half-spoof, the abandoned Mohole proposal of the 1950's, out of whose failure blossomed the ship's dynamical positioning scheme. That keeps the unmoored drilling vessel aligned to some 50 meters' accuracy above the drill hole kilometers below, trimming its position by sensing and responding to acoustic beacons set on the bottom.

It is no beauty, this slow-steaming ship of 11,000 tons, with a 60-meter drilling tower, a foredeck piled high with 10 kilometers of big steel drill pipes and a "moon pool" opened midships to pass the drill string and its bulky attachments. It can nonetheless support 500 tons of drill pipe, to bore into rock a kilometer below the ocean floor through a seven-kilometer layer of seawater. It has done something like that 500 times, in all the oceans except the Arctic, over its dozen years of service.

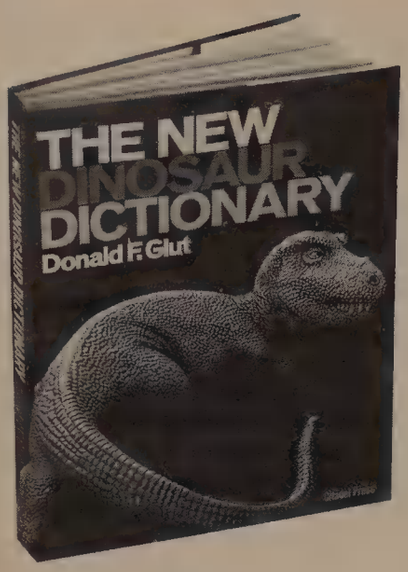
For two months in 1970 direct responsibility for 15 drill holes of its expensive mission (put the operating cost at $1 million per month at sea, and $100 million or so of 1968 dollars to build it) was shared by Hsü and Ryan, who had just finished his Ph.D. The wiseacres complained, with some show of reason, that the unique ship had been placed in the hands of "a student and an amateur." Those two co–chief scientists were, however, well up to their heavy task of supplying relevance to the ship's work, day after costly, tedious day.

The book is a personal and candid chronicle of Hsü's Leg 13, a tale of garbled radio messages, stubborn management, elated success, grim failure, shipboard intrigue and loyal and delighted colleagues. The cast of characters includes the captain and his crew, who knew both the sea and the demands of shore on the vessel; the drilling crew, roughnecks, tool pushers and their supervisors, who knew how to manhandle the rig and how to manage its surprising encounters with the rock far below; the half-dozen scientists who were "the 'decoding' section," and the long-haired young marine scientists and technicians from the Scripps Institution, who staffed the shipboard laboratories. The geophysical laboratory had to characterize and store the precious cores, ready for X-ray and mass-spectrometric analysis; the micropaleontological laboratory peered expertly at the tiny fossils that make possible stratigraphic dating, "the Foram-, Rad- and Nanno- team," as they signed their parting ditty.

There were 69 people aboard, "a big, happy, quarreling family in a small oasis, isolated from reality," insecurely led by what they found, their novice leaders acutely aware of the costs of their golden opportunities. Every time the drill stuck or the satellite position electronics (finding its first use outside the navies) went bad there began a new hour of tense decision. The lucky campaign of Leg 13 nonetheless sailed to success, to end in two months, rotating to shore its leaders and boarding others to man a new drilling cruise in another ocean.

One feature of the voyage is curious. The leg was nearly canceled by the planners because earlier seismic exploration of that sea floor had revealed abundant salt domes with strong promise of gas and oil. The industry was glad enough to have such a powerful exploration crew at work, particularly one it did not have to pay for. The National Science Foundation advisers, in an era of $3 oil and new ecological concerns, took quite the opposite view. A find would probably be too deep to be of any benefit, and it could pollute the seas. By imperative command from the heights above, the drill had to avoid broaching any submarine gas or oil. Worse than a sticking drill or a lost bottom assembly was the strong scent of gasoline in a core; get out of there!

There is now, of course, much more



When the first edition of **The Dinosaur Dictionary** appeared in 1972, it was named one of the best reference books of the year by the American Library Association.

In the years that have elapsed since then, there have been tremendous advances in the study of dinosaurs, making older books on the subject obsolete. In the words of Bob Long of the University of California's paleontology department, we are entering a new Golden Age of Dinosaurs. More new dinosaurs are being discovered than at any time since the first Golden Age back in the nineteenth century, and there have been many reclassifications. Some specimens once thought to be dinosaurs have now been shorn of that designation.

Donald F. Glut, now a member of the Society of Vertebrate Paleontology, based the first edition of this book entirely on his own careful research of the field. As a result of its success, he has been deluged with advice and new information from all over the world — from the best "dinosaur men" in the field.

The New Dinosaur Dictionary is exactly what its name implies. It is an alphabetical listing of every genus of dinosaur now known to paleontologists. It contains more than six hundred illustrations — drawings, paintings and photographs of skeletons and reconstructions. Many of the pictures have been prepared especially for this book by well-known dinosaur illustrators, including 25 by William Stout.

It is by far the most comprehensive and up-to-date book on the subject of dinosaurs.

Published by Citadel Press

$19.95 at booksellers. Or you may order your copy directly from the publisher by using the handy coupon below. Do it today!

CITADEL PRESS* SA
120 Enterprise Avenue, Secaucus, N.J. 07094

I enclose my check or money order for $21.95 ($19.95 plus $2 for prompt United Parcel Service delivery). Please rush a copy of THE NEW DINOSAUR DICTIONARY by Donald F. Glut to

My name: _____
(please print)

Street: _____

City: _____ State: _____ Zip: _____

***Citadel Press is a Lyle Stuart company.**

evidence than those early flakes of gypsum. The cores brought up samples of algal stromatolites, signs of shallow-water growth in the sun, and of "chicken-wire anhydrite," a wispy distribution of dark carbonates in a white matrix of calcium sulfate, long taken as a signature of arid coastal flats in the borehole studies of petroleum geologists. By now the carefully archived half of core 124, left undisturbed in the first shipboard studies, has been painstakingly peeled. Eight or 10 cycles can be traced, one above another; the laminations follow in regular succession, first the fines of a lake bottom, signaled as brackish by the unusual diatoms found, then layers more disturbed as the lake shallowed and the bottom felt the winds, then the surface algal growth, then the final aridity shown by the chicken-wire anhydrite. The cycle repeated as seawater entered anew or the next lake overflowed; for a million years or so that was the history of a random spot on the Great Mediterranean Desert.

The case is hard to doubt. All plausible alternatives stand excluded. The micropaleontologist Maria Cita of the University of Milan, one of the shipboard specialists to support the early insights of the cruise, has by now helped to piece together the chronology of the salinity crisis of that time, when the marine organisms almost vanished from the cores. Today whatever evaporating seawater there gets briny and hence denser can escape out into the Atlantic, flowing deep below the surface of the Strait of Gibraltar. In the Miocene, however, Gibraltar was a huge dam between the Mediterranean and the Atlantic. The waters got steadily saltier in the hot sun; only a dwarf microfauna survived. Suddenly the dam broke and the great basin was refilled over a falls a thousand times grander than Niagara, taking perhaps centuries.

The microfauna of the modern Mediterranean does not resemble its Miocene predecessors; it is more like the Atlantic fauna. The animals are immigrants, not natives. The old plankton families are gone, lost in the salt. The Rhone and the Nile cut deep gorges into the dry valley, long drowned; there are many others. The caves and the karst lands of the coasts are now understandable; they were highlands draining into a low desert. In 1975 the *Glomar Challenger* returned to drill again deep below Mare Nostrum. This time the cruise was "more somber and more difficult." The cores probed deeper and tapped earlier times, revealing a deep-sea environment for at least 15 million years before the desert epoch. The catastrophe of desiccation has left its residue of rich potash deposit, the last salts to crystallize out of the brine, just where they ought to be, in the deep Ionian basin. Then came a series of lakes fed by the reorganized streams of Europe, captured out of the north by that vast yawning valley.

Not everyone is convinced. Like the glaciers, this is a story hard to accept. Now, however, we take the ice for granted. "Cita gets excited sometimes when her colleagues do not listen to her." One of these days "Cita's last tormenter" will fall quiet, and the desert of the Mediterranean will be held as gospel truth. Then in the megayear fullness of time, if our species somehow makes it, it will all happen again. The Strait is shoaling up as the plates compress the basin. Once again there will be an isthmus at Gibraltar. Cannes will become a highlands resort for the desert roughnecks winning the oil from all those salt domes southward, and the more adventuresome tourists will price dory trips down the twisting gorge of the Nile and camel caravans to the mountains of Malta.

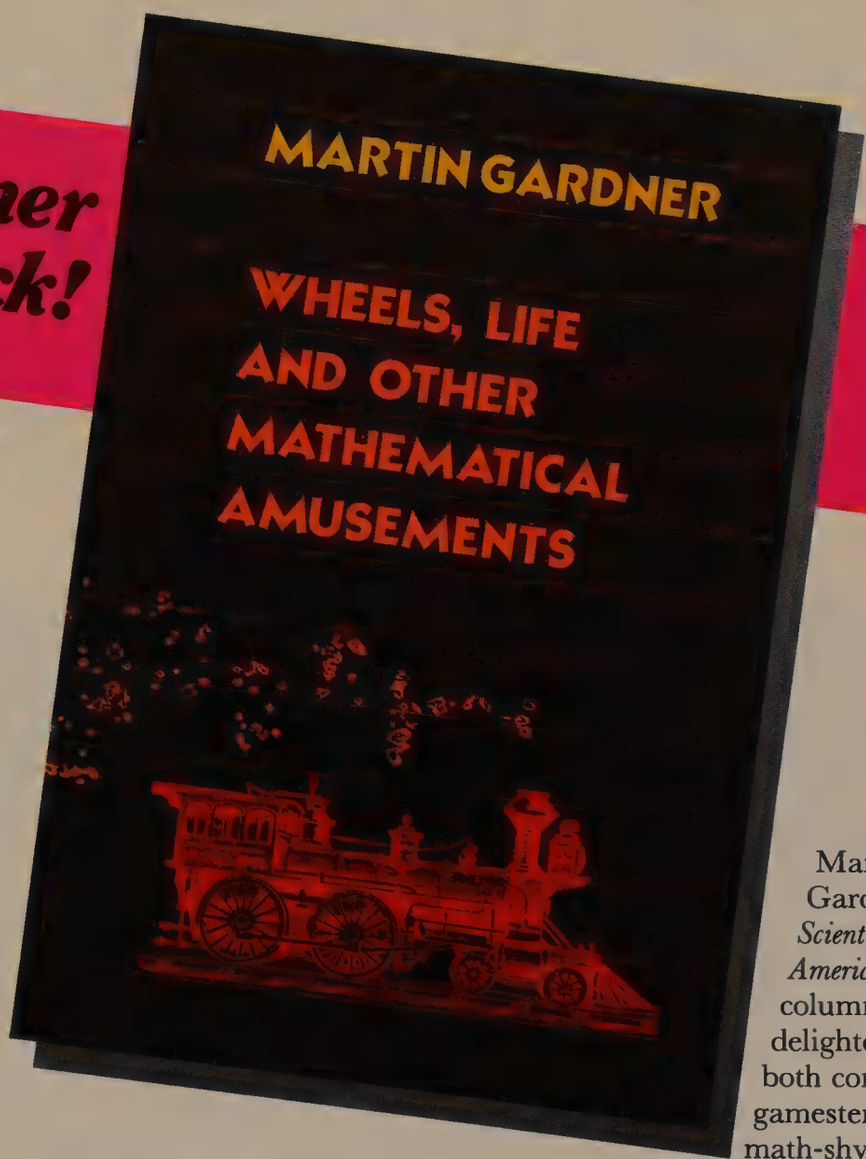
Kenneth Hsü is professor of geology at the Swiss Federal Institute of Technology now, and no amateur. But his warmth, good humor and candid manner remain much as they must have been when he spent all those days and nights forward in the noisy little driller's shack just starboard of the tower. The plentiful photographs, maps and bottom profiles are evocative and helpful. This is a winning account of a sunny and serendipitous quest after deep secrets, of youth and insight and simplicity amid worldly tangles, and not least a "testimonial to a friendship born at sea."

THE MATHEMATICAL SCIENCE OF CHRISTOPHER WREN, by J. A. Bennett. Cambridge University Press ($29.95). DEVELOPMENTS IN STRUCTURAL FORM, by Rowland J. Mainstone. The M.I.T. Press ($17.50). "Since the time of Archimedes, there scarce ever met in one man, in so great a perfection, such a Mechanical Hand, and so Philosophical a Mind." So wrote Robert Hooke of his friend Christopher Wren, in terms apt for Hooke himself, both men in their early thirties. Within a few years Wren had become Surveyor-General of the Board of Works; his London had burned, and it was his responsibility to build it anew, including that most unselfish monument to the architect himself, St. Paul's Cathedral.

J. A. Bennett has made a case in his slender book that Wren's brilliant career as mathematician, instrument designer and astronomer was not twisted awry on his entry into public architecture but was fulfilled. Those 50 parish churches and Wren's own Greenwich Observatory today delight the London visitor as much as the sight of his great dome, itself a little diminished by the careless power of our century. Indeed, his practical bent and his delight in proportionate form were foreshadowed in his mathematics. His aesthetics, like his science, derived not from higher metaphysical

Wheels, Life, and Other Mathematical Amusements

Martin Gardner is back!



"In WHEELS the prestidigitating master is better than ever."
—Isaac Asimov

Four years have passed since the last volume in this highly successful series drawn from Martin Gardner's popular "Mathematical Games" column in *Scientific American*. This new collection of problems and paradoxes, teasers and tricks is guaranteed to enhance his already considerable reputation as a master conjurer.

Ranging from classical paradoxes to geometric fallacies and games, WHEELS, LIFE, AND OTHER MATHEMATICAL AMUSEMENTS includes entertainments as easy as pi to wonderfully challenging excursions into the intellectual beyond. With anecdotes about famous (and not-so-famous) mathematicians sprinkled among the games and problems, WHEELS, LIFE, AND OTHER MATHEMATICAL AMUSEMENTS will provide hours of enjoyment for recreational math devotees, games buffs, or those who simply want to sharpen their wits.

Martin Gardner's *Scientific American* column delighted both confirmed gamesters and math-shy novices for almost 20 years. This new treasure-trove is sure to win him even more fans.

350 pages, 172 illustrations; $15.95 (cloth)
ISBN 0-7167-1588-0

For your copy, fill out the accompanying order form, or send payment (including $1.50 for shipping and handling; California, New York, and Utah residents add appropriate sales tax) to:



Order Department, W. H. Freeman and Company, 4419 West 1980 South, Salt Lake City, Utah 84104.

principles but from a very English recognition of sensed geometric fitness. He followed Palladian models all right, but he would not obey rules that were "too strict and pedantick." He held a scrupulous regard for the relation of form and use. "Churches... must be large," he wrote, "but still, in our reformed Religion it would seem vain to make a Parish-church larger, than that all who are present can both hear and see. The Romanists, indeed, may build larger Churches, it is enough if they hear the Murmur of the Mass... but ours are to be fitted for Auditories." And we still enjoy his Observatory in the Park, with its facade evoking some grand musical instrument, built at almost domestic scale for use and convenience yet leaving "a little for Pomp."

Rowland Mainstone is an expert on contemporary building design and construction, a senior staff member of the British Building Research Establishment. He admits to, indeed he displays, a long interest in earlier counterparts of our present problems and solutions, which he has "never found it easy to regard... as something apart." In this large paperback (it is only a reviewer's oversight that has delayed the notice of this admirable book until its bargain edition) his thoroughly analytic and entirely nonmathematical text is itself built on a splendid plinth of photographs and drawings, to show us hockey rinks and temples, coal bunkers and pyramids, mosques and the TWA terminal at Kennedy. The views often include apposite detail; Mainstone is a widely traveled consultant (although not to the Far East) with a talent for photography.

The book is a history not of architecture itself but of the growth of structural form, of course bound to the materials in which it has been realized. Its organization is not chronological but topical, a scheme much better suited to Mainstone's reflective look at building. The first part includes a treatment of the requisite background study in the mechanics of equilibrium and the response of building materials. The second part discusses the elements of structure one by one, not only vaults, beams, trusses and space frames as they have arisen but also a particularly welcome chapter on their joining and their supports by walls and foundations, easily overlooked but literally the basis of all construction. A final part then treats of structures as a whole: buildings, bridges and towers old and new.

History is sampled for its clues to structure. The architecture of old Athens and that of late Renaissance Europe are important less for new structure than for the way they "exploited existing structural possibilities." Hence they are here given less attention than the days of the Roman emperors, the French Gothic and our own times, when structural advances were the top of the news. The result is a treasure for readers with and without some understanding of how buildings stand, a first-rate new entry in the short list of admirable books on the fight with gravity, one that spans the history of what and why we have built with a sturdy and persuasive account of how.

Look into a few of the 16 chapters. Building is never much like growth, except for those remarkable caves, temples and shelters carved out of the living rock, and perhaps for inflated structures. It is discontinuous; intermediate stages either are incomplete or cannot stand alone. Even defining the designed form at large scale, easy enough perhaps in a plane on the ground, requires something more to set it out high in the air. The prefabrication of masonry is as old as the Great Pyramid itself, and the Romans set up arches in their aqueducts first stone by stone horizontally on the ground; the marks for the layout can still be seen.

The deck truss of the new suspension bridge across the Firth of Forth is shown here in construction, its free ends curving sharply upward before it feels the full load of the long span under cable



support. Prestressing has a history as old as ship's rigging and wagon wheels with iron tires; its use today waited for better understanding and better jacks. Both an external prestress, say on the steel reinforcing cables in concrete beams, which is applied while the member is quite free, and a prestress applied to the finished element in place, in effect squeezing or stretching it to fit, are in common service, ever since the pioneer concrete designer Eugène Freyssinet 50 years ago and more.

A stepped and timber-linked piling holds up the Rialto bridge across the Grand Canal, quite new in the 16th century. The complex built-up struts of the railway cantilevers across the Firth of Forth make plain how much effort went into that wonderful Victorian production, whose legible three-dimensional form—the photograph is a delight—demanded an intricacy of detail hardly acceptable in these times of fairer wages and faster, safer work. Then there is Wren himself; his design for the dome of St. Paul's was one of the earliest known to embody mechanical calculation, to be sure in a rather misleading Archimedean approximation. He neglected the horizontal thrust between the two conceptual halves of the dome, assuming stability if each center of gravity lay above the weight-bearing pier. The dome stands very well indeed; prudence, experience and intuition are great aids to bad theory. The first acceptable rational analysis of a structure of which we know is the report on the dome of St. Peter's in Rome made by three very able mathematicians in 1743, T. Le Seur, F. Jacquier and the famous R. G. Boscovich. The form of the alarming new cracks then visible at the springing of the vault certainly steered them away from Wren's error; they put on ties against outward collapse, which have worked well ever since, "epoch-making," the author adds drily. Even though it is not quite clear that they understood fully, their analysis was correct.

THE HUMAN TERATOMAS: EXPERIMENTAL AND CLINCAL BIOLOGY, edited by Ivan Damjanov, Barbara B. Knowles and Davor Solter. Humana Press, P.O. Box 2148, Clifton, N.J. 07015 ($49.50). Every reader would accept the suggestion that cancer is protean, manifold. Yet for most of us there is a simple model. Somehow one cell of a normal tissue eludes the developmental controls. It divides into a clone of similar progeny that may come to abandon the specialized site and even the nature of their ancestor, to spread as an invasive cellular army of mutineers.

This up-to-date although technical survey by a passel of experts from Europe and America serves first to show how schematic that model is. A teratoma (the word is from the Greek for monster, plus the ending familiar for neoplasms) is a tumor composed of multiple well-differentiated tissues, profoundly foreign to the site where they appear. The commonest form in children presents at birth, as "a grotesque protrusion" from the low end of the spinal column, mainly in female infants, perhaps in one birth out of 26,000. That diagnosis requires no opening of the body cavity; it was described long before Galen, in clay tablets from Nineveh, as the birth of an infant with three feet, and an omen of prosperity. Nowadays such tumors are easily removed and are generally benign. They contain a variety of mature tissues, in particular skin, nerve and gland.

Commoner and more dangerous are teratomas of the ovary and the testis, appearing usually in young adults. The ovarian type, both disclosed and removed by surgery, are most commonly dermoid cysts, sacs the size of a tennis ball or larger—the record is beach-ball size—that turn out to hold an inner layer of well-developed skin, with its details of sweat and sebaceous glands and a fatty underlayer, the skin usually overgrown with dense hair. A third of them





THE PERSONAL PRINTER BUILT LIKE A SHERMAN TANK PERFORMS LIKE A STRADIVARIUS.

A Tough Act To Follow. Frankly, an Okidata printer is the best printer on the market today. Why? We pack more performance features per dollar into our dot matrix printers than anybody. Bar none. Data processing at speeds from 80 to an incredible 350 characters per second. Exceptional letter quality printing at three times the speed of most daisywheels. Draft and color printing. High resolution all points addressable graphics for charts, graphs, illustrations and photos. Even alternate character sets for self-designed typefaces and symbols. (Print a G-clef, if you like.)

Long-Term Engagement. Our high performers are the most reliable printers available, with rugged steel frames, laser-welded parts, and a print head (the most vital part of all) so strong we guarantee it for up to one full year. No wonder our warranty claim rate leads the industry at less than ½%.

Perfect Harmony With Your Computer. Every Okidata printer works in concert with the major names in personal computers. Better and faster, in fact, than the major computer name printers. (Which, by the way, aren't even made by the major computer companies.)

Best Selection. Right Price. Nobody offers you a better choice. There are eight Okidata printers in all, ranging in price from $449 to $2995, suggested retail. Suggest you call 1-800-OKIDATA (in NJ, 609-235-2600) to find the dealer nearest you. You'll get more for your money from your personal printer. Not to mention your personal computer.



OKIDATA

Mt. Laurel, NJ 08054

A subsidiary of Oki Electric Industry Company Ltd.

contain teeth as well; the record tumor held 300! Thyroid and pancreas tissue are not uncommon.

The testicular analogues of the ovarian tumors are usually less differentiated, fast-dividing, immature, embryonic and thus much more dangerous. They never contain teeth. One teratoma in 10 or 20 arises at sites outside the gonads. Strangely enough, these tend to occupy the midline of the body, all the way from the lower pelvic region to the rear lining of the abdomen or the membrane dividing the right lung from the left, up to the pineal gland within the cranium: symmetry in disorder. The most complex structures repeatedly found are fingers, complete with bones, muscles, fat, nerves, skin and nails. A couple of documented cases of more or less fetuslike tumors are cited.

The various chapters of this comprehensive review deploy the entire battery of modern techniques and viewpoints. In the clinical direction the search is on for specific proteins secreted by these tumors, in the hope of monitoring the tumor burden in patients under chemotherapy, and even for diagnosis. Such a watch depends on the use of antibodies linked to fluorescent stains; here the work parallels that on more ordinary tumors. A careful course of combined chemotherapeutic drugs has quite recently "made testicular cancer one of the most curable malignancies," with a cure rate expected to be 95 percent of the 4,000 cases per year in the U.S.

A couple of papers recount the progress of tissue cultures of cell lines taken from human teratomas. Cultures allow much more detailed cytology than would otherwise be possible, although the work is rather new and has not yet yielded cells that differentiate very strikingly. Genetic studies of neutral biochemical marker genes in host and tumor cells have given good reason, however, to believe the dermoid cyst arises from a single cell. Its DNA has replicated and chromosomes have been paired, but it has become a diploid without fertilization, parthenogenetically. No such detail can be adduced for the origin of any other tumor.

The laboratory mouse has provided even deeper insight by way of these strange disorders. Some 10 years ago an inbred strain of mouse was found in which about half of the animals developed spontaneous ovarian teratomas. At once the earliest steps in development were opened to study; before that time only fully developed tumors were much known. There has been fast progress; by 1981 it was shown that selected cells of the mouse teratoma—stem cells of distinct appearance—could differentiate and give rise to many tissues in a line of normal mice if some normal embryonic cells also were available for cooperation. So far it has not been possible to generate normal mice solely from clumps of such teratoma-derived cells. Perhaps that will yet work. The old idea is at the point of demonstration: differentiation is the antagonist of proliferation. The tumor stem cell, like the primordial germ cells of each mammal, carries the information to become any phenotype of the animal, but it loses that chance as it differentiates along any single path of its potentialities. When such a cell escapes an early embryo, it may begin to form a teratoma.

Human beings are not mice. There are plenty of distinctions among their teratomas too, but it looks as though this wry direction of research points toward the large questions of development, a chain of alternative gene expressions, a program longer and subtler than can well be extrapolated from bacteria. Teratomas are defects of development of a logical hierarchy quite different from that of tumors in general. Pathology once again illuminates, if still fitfully, the path to understanding those extraordinary phenomena we call normal.

RIEGEL'S HANDBOOK OF INDUSTRIAL CHEMISTRY. Eighth edition edited by James A. Kent. Van Nostrand Reinhold Company, Inc. ($59.50). "As always, the aim of this book is to present an up-to-date account of the many facets of as broad a cross section of the chemical process industry as is reasonable in a single volume." So writes the editor, a chemical engineer, now dean at the University of Detroit, who has gathered 34 authors, most of them experts from within the industry, for his account. They have prepared some 20 chapters, each one devoted to the industry of a class of related products, from sulfuric acid to sugar, from oils and fats to explosives, from rubber to dyestuffs, from pharmaceuticals to sweeteners. A few chapters take the standpoint of input: coal technology, petroleum and synthetic organics. There is a still more diverse chapter on the nuclear industry, from ore and reactors to stable isotopes. Two other general chapters treat economic aspects and the technology of wastewater.

The book is the inheritor of a classic tradition; the first edition appeared from Riegel's own hand about 50 years ago. In this edition the material can be dated to about 1980. The heavy volume offers a readable and concise engineering summary of a complex of industrial activities; a needed chapter on the labor force and another on the supplies of process equipment are both lacking. Some of the flavor of an excursion through these nearly 1,000 pages can be transferred by an assortment of examples of recent swift change, whether in a process, a raw material or a marketplace.

Sulfuric acid is the workhorse of the old-line chemical industries. Its catalyt-

BOOKS

Tatsch Associates specializes in research and consultation about the Earth's energy and mineral resources on a global scale. The research results are summarized in the following hardcover books, which are available directly from us or through book dealers in all parts of the world.

Coal Deposits: Origin, Evolution, and Present Characteristics. 1980. 590 p., illustrations, bibliography, index, $156.00

Copper Deposits: Origin, Evolution, and Present Characteristics. 1975. 339 p., illustrations, bibliography, index $72.00.

Earthquakes: Cause, Prediction, and Control. 1977. 451 p., illustrations, bibliography, index. $108.00.

The Earth's Tectonosphere: Its Past Development and Present Behavior, 2nd ed. 1977. 468 p., illustrations, bibliography, index, $30.00.

Geothermal Deposits: Origin, Evolution, and Present Characteristics. 1975. 275 p., illustrations, bibliography, index, $84.00.

Gold Deposits: Origin, Evolution, and Present Characteristics. 1975. 275 p., illustrations, bibliography, index, $72.00.

Iron Deposits: Origin, Evolution, and Present Characteristics. In prep. 600 p., illustrations, bibliography, index, $180.00 prepublication; $216 thereafter.

Mineral Deposits: Origin, Evolution, and Present Characteristics. 1973. 264 p., illustrations, bibliography, index, $64.00.

The Moon: Its Past Development and Present Behavior. 1974. 333 p., illustrations, bibliography, index, $20.00.

Petroleum Deposits: Origin, Evolution, and Present Characteristics. 1974. 378 p., illustrations, bibliography, index. $50.00.

Uranium Deposits: Origin, Evolution, and Present Characteristics. 1976. 303 p., illustrations, bibliography, index. $96.00.

TATSCH ASSOCIATES
120 Thunder Road
Sudbury, Massachusetts, 01776, U.S.A.



Journal of Field Archaeology

Taxonomists

in all branches of botany, zoology, and medicine are needed to identify archaeological finds. The *Journal of Field Archaeology* is preparing a list of taxonomic specialists who are able to classify plant (seeds, pollen, fibers, wood, etc.), animal (insects, shells, bones, hair, teeth, leather, etc.), and human remains from current archaeological excavations. If you are willing to participate in such inter-disciplinary work, please send your name, address and the materials you are competent to identify to:

Curt W. Beck
Vassar College, Box 92
Poughkeepsie, N. Y. 12601

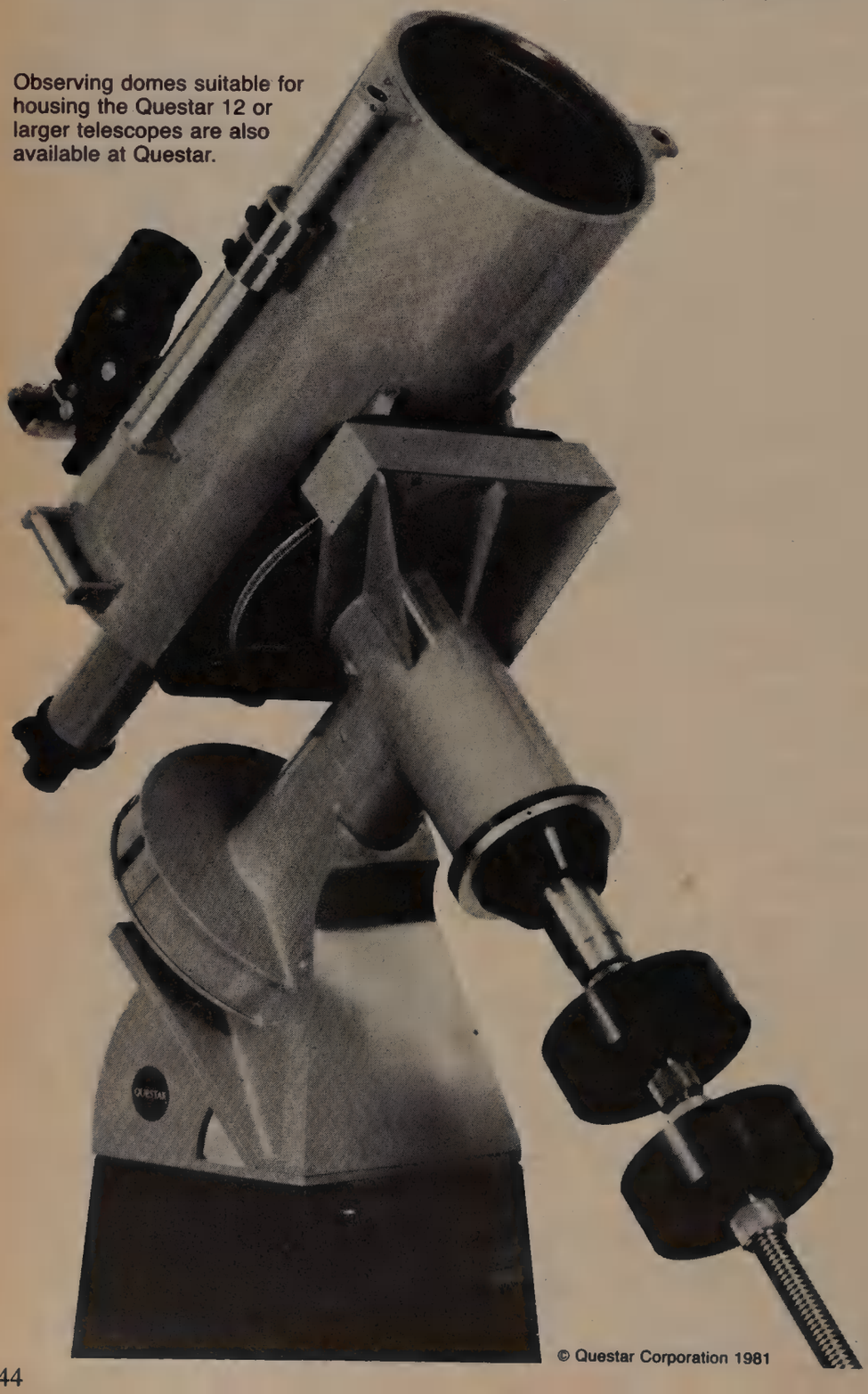
QUESTAR® 12 on the QUESTAR® MOUNT

The Questar 12, latest addition to the Questar family of fine optical instruments, now has its own Questar-designed mount. A German equatorial type, it is notable for its 360° continuous tracking in R.A. with precision tracking to better than 4 arc seconds. The Questar Mount is designed with over-sized components so that it can accommodate any Questar up to 18 inches. The standard mount shown is straightforward in design but can be modified so as to be compatible with more sophisticated tracking devices or other special equipment.

The Questar 12 is a superb instrument for the serious astronomer, for the university astronomy department or the engineer seeking sophisticated tracking and surveillance equipment for which Questar Corporation has a noted reputation.

Questar Corporation Box C, Dept. 20, New Hope, Pa. 18938 (215) 862-5277

Observing domes suitable for housing the Questar 12 or larger telescopes are also available at Questar.



© Questar Corporation 1981

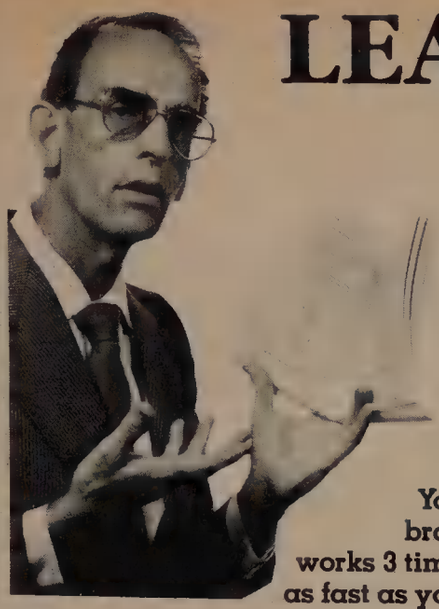
ic production is so rationalized by now that plant location is set by the location of the market, since the weight of sulfur is small compared with the weight of the acid, made up by air and water. This industrial symbol is produced in no state half as much as it is in rural Florida! The reason is that the phosphate rock mined in central Florida requires the acid to yield superphosphate fertilizer. The sulfur used in the U.S. these days is mostly extracted very pure from the sour natural gas of western Canada; the sulfur from drilled wells has lost its market lead.

Ammonia is the chief form of nitrogenous fertilizer today; it has gained greatly in the U.S. over 20 years, and construction is under way on many large new plants in China, India and the U.S.S.R. Asia (outside the U.S.S.R.) now synthesizes more ammonia than either of the superpowers. Sometimes, however, nature catches up with the chemical engineer. In the U.S. sodium carbonate—the indispensable ingredient of glass—is no longer supplied by the Solvay process, the classic mark of heavy industry, with its waste pools of coal-blackened calcium chloride solution. Long, long ago the volcanoes made sodium carbonate in bulk out in Wyoming; the curve of production for Solvay synthetic soda ash has fallen steeply, to drop below that of the mined ash 10 years back, and the mines are still gaining. Naval stores—rosin and turpentine mainly—were once the distillate of the stumps and roots of the southern pine, or of the gum collected from the living tree. Now nearly all of them arise as by-products of the kraft-paper industry, as vapor out of the top of the big continuous digesters odorously making strong dark pulp out of Southern pinewood chips.

Technology can now spread the bacterial enzyme glucose isomerase out along a big vertical column. Starch from any source, say corn-mill slurry, can be hydrolyzed to a glucose syrup with acid, and the glucose can be converted into the much sweeter sugar fructose by a pass through that active column. Commercial use of the synthesized sweetener is mushrooming, although it is still no overall threat to sugar. Oxygen is now a megatonnage chemical (nitrogen and argon too); pipelines distribute the pure stuff in gaseous form for a few miles around a central air liquefier to the steel ingot furnaces and chemical plants that are the big consumers, for example at Gary and Houston.

The chapters vary in readability; all seem authoritative, although the sincere authors do not always maintain detachment from the problems they see besetting their own effluents and price structure. Safety and externalities have not yet found a productive place in every supply-side mind.

LEARN MATH — *EASY!*



Your brain works 3 times as fast as your fingers can write! That's where the Calculator comes in — you can go three times as fast with your machine doing the dirty work.

It's easier and more interesting. And you'll understand it better when it's fun. (If you haven't yet tried out a modern scientific-statistical calculator, you're in for a pleasant surprise — see below!)

WHY MATH? To get a promotion or a new job. Just to hang on to that job you've got. To quicken your mind. To understand this world you live in. To impress people. Just for some fun you'll feel good about. (Read those letters below to see how it has worked for others.)

Do it at home... You can start right away and go as fast as you like! Put it aside while you're busy with something else — in short, *set your own pace!*

Lots of us like to avoid being compared with the other members of a class — we like to set our own standards. Too, we like to save the time and expense of going to a school or seminar. (Best of all, it's a lot cheaper to study at home where the rent is already paid — *read on!*)

Let's be honest. You wouldn't expect to learn to play baseball or chess by watching an expert do it. Math is the same — you'll need to do some work. BUT — with my calculator method, you'll see it's a whole new ball game!

THIS METHOD IS FRIENDLY! You can use it on your own. You follow interesting examples step-by-step with calculator keystrokes — instead of algebra.

Let me introduce myself — I'm Dr. George McCarty and I teach at the University of California. Tens of thousands of people like yourself have already used my guidebook to teach themselves calculus. Recently I have teamed up with another professor. We put those same tested and proven techniques to work for you in a new *Statistics Guidebook.*

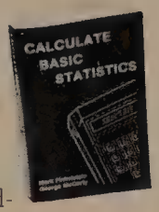
Even if you've had trouble with math, this method will work for you. Choose the one you want to learn, calculus or statistics (or both) and start *today!*

As pennywise Ben Franklin said, "An investment in knowledge pays the best dividends."

CHOOSE STATISTICS

It's used everywhere — to set insurance rates and to assess a drug's effectiveness, to detect a crooked roulette wheel and to measure lightbulb lifetimes, to audit the books and to find your position at sea.

People have always known that their experience was not a perfect guide to the future. Statistics shows us that uncertainty is not only inevitable but it itself is measurable.



You'll find out about this in my guidebook. You'll read informal explanations with detailed calculations for topics like "confidence intervals" and "contingency tables."

YOU'LL BE DOING Inference and Hypothesis Testing, the Chi-Square Distribution, ANOVA, Regression and Correlation, Non-Parametric Methods — they're all here.

SUPPOSE YOU HAVE ALREADY STUDIED STATISTICS. You'll use Quick-Reference Guides to go straight to a topic you want to refresh yourself on — then get *right-now* answers in your own field, business, biology, science, psychology, education.

The Guidebook is guaranteed to work with any calculator or computer. If you would like to start out with a fancy new machine, order my Statistics Kit which includes the guidebook and a TI-35 calculator which finds means and standard deviations at the touch of a button. It has an LCD display, 1000 hour batteries — the works.



PHONE MC or VISA ORDERS TO
(714) 831-2637
Ask for our Catalog of books and calculators.

CHOOSE CALCULUS

The math of science and engineering — satellite orbits and compound interest, bacterial growth and radioactive decay, marginal profits and reaction rates.



You'll find out about these concepts when you use the Calculus Kit. This Kit contains my guidebook and a TI-35 calculator which finds log and trig values at the touch of a button. (It's modern, too, with LCD display and 1000 hour batteries.)

In the guidebook you'll learn both differential calculus and integral calculus — *even differential equations!*

NEED TO EVALUATE functions, areas, volumes — solve equations — use curves, trig, polar coordinates — find limits for sequences and series? It's all here!

If you're in the biological, social or physical sciences, you'll be doing Bessel functions, carbon dating, Gompertz' growth curves, half-life, future value, marginal costs, motion, cooling, probability, pressure — and plenty more.

Important numerical algorithms are here, too: rational and Padé approximation, bracketing, continued fractions, Euler's method, Heun's method, iteration functions, Newton's method, predictor-corrector, successive substitutions, Simpson's method and synthetic division.

If you already have a scientific calculator, you can get just the *Calculus Guidebook.*

George McCarty

P.S. — YOU *CAN* DO IT! I'm certain I can show you this easy way to success with math, so ***I guarantee it to you unconditionally*** — if you are not satisfied just send it back for a prompt refund!

LOOK AT WHAT THEY SAY

- "...*excellent diagrams and clear explanations.*" —POPULAR SCIENCE
- Professor John Ball of Harvard College writes: "*I wish I had had as good a calculus course.*"
- S.C. McCluney, Jr. of Philadelphia writes: "*With your book and a calculator the whole idea becomes clear in a moment, and is a MOST REFRESHING EXPERIENCE.*"
- Professor H.I. Freedman of the U. of Alberta, states, "*There can be no question as to the usefulness of this book...lots of exercises...very clearly written...easy reading.*"
- "...*straight-to-the-point explanations...*" —HP Key Notes
- Tektronix Engineer Bill Templeton says, "*The best, most clearly written book I have seen for improving your math skills.*"

TAX DEDUCTIBLE FOR PROFESSIONALS

YES, I want to learn math **fast** and **easy** with your *new method.* If I am not more than satisfied I can return it for a quick refund. RUSH me the guidebook or complete kit I select below (kit contains the guidebook and a modern TI-35 calculator).

☐ Send **FREE** Catalog of books & calculators

STATISTICS GUIDEBOOK		KIT
How many?	☐ $14.95	☐ $39.95
Handling/Shipping	1.00	2.00
Tax (within Calif.)	.90	2.40
TOTAL		

CALCULUS GUIDEBOOK		KIT
How many?	☐ $19.95	☐ $44.95
Handling/Shipping	1.00	2.00
Tax (within Calif.)	1.20	2.70
TOTAL		

For AIR SHIPMENT add $3.00 per item

Phone (714)831-2637 (MC or VISA only)

Enclosed is ☐ Check, ☐ M.O.; or charge my MC/VISA... exp._____

Card #_____

Name _____

Address _____

City_____

State, Zip _____

***OUTSIDE USA,* surface shipping is $2.00 per book or $5.00 per kit. Air shipment is $5.00 per book or $10.00 per kit. U.S. FUNDS PLEASE.**

EduCALC Publications, Dept. A8
27953 Cabot Road
South Laguna, CA 92677

© 1982

Established 1845 SCIENTIFIC AMERICAN September 1983 Volume 249 Number 3

The Dynamic Earth

Presenting an issue on the earth as a system of interacting fluids, including living matter. Some of the flows are fast, others slow, but overall the planet is maintained in a remarkably steady state

by Raymond Siever

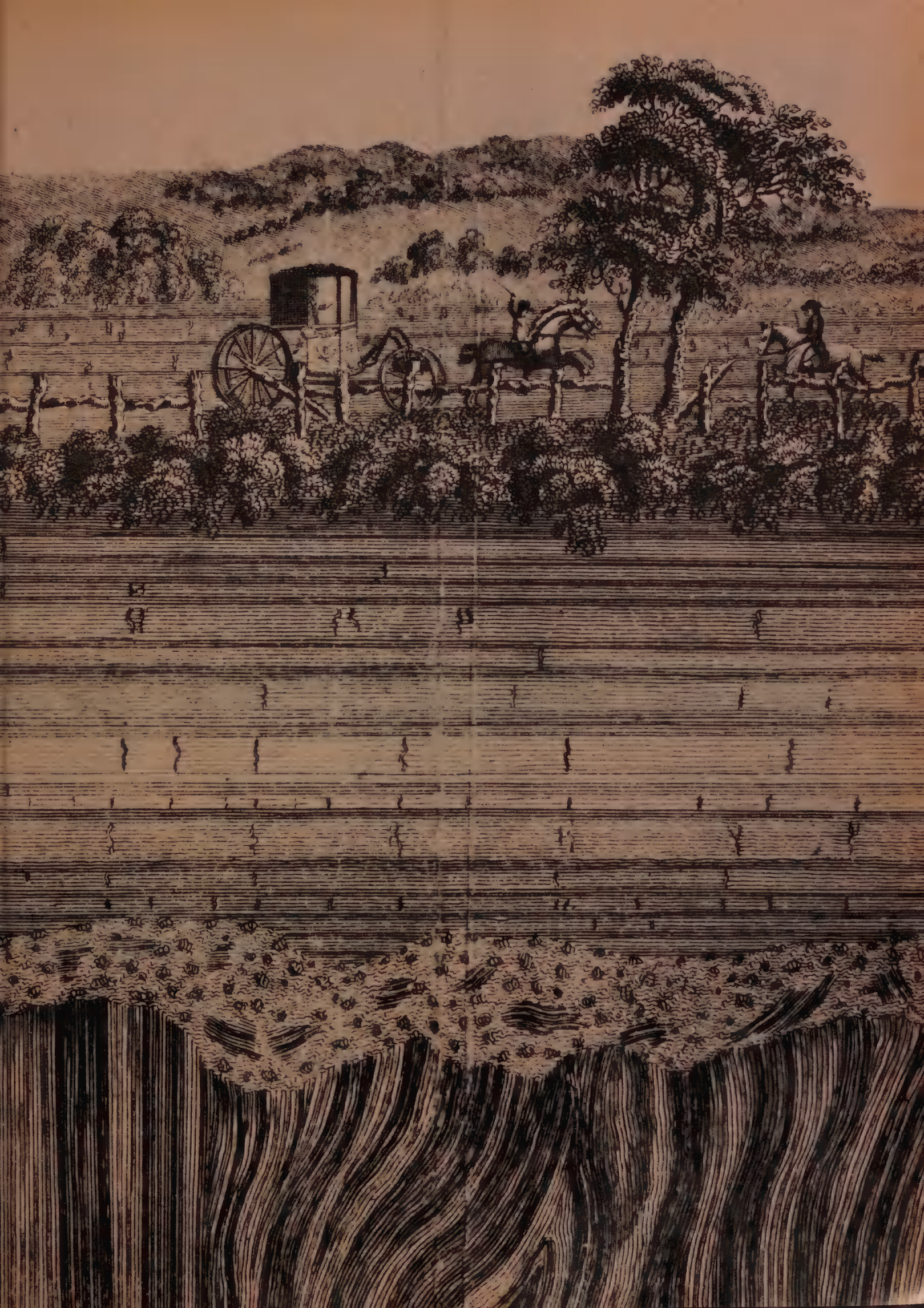
Scientists who study the earth are accustomed to working with many different scales of space and time. The physical dimensions of their subject range from the global to the submicroscopic, from volumes of matter measured in cubic kilometers to interatomic gaps measured in angstrom units. Often a given research topic will encompass widely disparate scales of length, as when an earthquake triggered by a slippage of a few centimeters along a fault generates seismic waves that travel for thousands of kilometers through the earth's interior. Similarly, the temporal dimensions of geology include not only short-lived phenomena such as earthquakes, volcanic eruptions and meteorite impacts but also events recorded in units of tens or hundreds of years (for example the meandering of a river), thousands of years (glaciation), millions of years (continental drift) and even billions of years (the formation of the present, oxygen-rich atmosphere). Again a single process, notably weathering, can be studied over a wide range of time scales: from the minutes and hours of a laboratory experiment measuring the dissolution rate of a mineral to the thousands of years needed to form a soil. Taken in various combinations, the parameters of geologic space and time define the scope of this single-topic issue of *Scientific American:* the multitude of great and small changes that have taken place—and continue to take place—in the history of the earth.

Most geologists, oceanographers and other earth scientists tend at one time or another to think of the earth as a machine, or perhaps as a living organism. The metaphor of the machine captures an important aspect of the earth's dynamism: in spite of all the changes that are observed at many different scales of space and time, the earth as a whole stays remarkably constant. In particular, it has become apparent in recent years that major parts of the globe—the core, the mantle, the crust, the oceans and the atmosphere—can be profitably viewed as a complex, interacting system in which there is a cyclic flow of material from one reservoir to another. The mechanistic model of the earth as a vast recycling system has its counterpart in the physiological model of dynamic equilibrium known as homeostasis.

The hierarchy of scales that pervades the work of the earth scientist is perhaps best exemplified in the making of a geologic map, an artifact that can be described in somewhat nongeologic terms as a graph of the position on a coordinate system at the earth's surface of rock formations of different ages. The first step in geologic mapping is the job of the field investigator, who determines two major properties of the rocks at a given site: age and composition. At a typical rock outcrop one can see only small-scale relations, usually over distances measured in meters. From an assemblage of such observations the final map of the region is constructed, like any other graph, by interpolation and extrapolation, showing forms appropriate to its scale. On a map with an area of, say, 200 square kilometers one can see patterns of river valleys and the characteristic folds and faults of bedrock. The wealth of information at the outcrop level is sacrificed in favor of grosser features. On a map that covers a region of many thousands of square kilometers one begins to see even larger features: plateaus, mountains, plains, entire river systems, the outlines of a rift valley, the distribution of glacial lakes. As the map gets to continental or global proportions one sees the largest structures of the continental surface, principally mountain chains. Whenever detailed information is sacrificed for patterns that show up only on a large-scale map, the trick is to recognize what details to leave out. In other words, the essence of this kind of geologic analysis is always to separate the "signal" from the "noise."

Inevitably earth scientists face the problem of reconciling scales. For example, structural geologists and geophysicists are currently trying to relate the large-scale collisions of tectonic plates that throw up high mountains such as the Alps and the Himalayas to the small-scale folds and faults that can be seen on an individual mountainside. The aim is to learn how to go in the opposite direction: to infer from small-

EVIDENCE OF ANCIENT UPHEAVALS underlies the pastoral scene on the opposite page. The engraving is from James Hutton's *Theory of the Earth,* published in 1795. It shows an exposed riverbank of the river Jed in southern Scotland. The vertical beds of rock at the bottom of the bank were originally laid down as oceanic sediments. They subsequently underwent metamorphosis to become schist and were deformed and uplifted to become part of a mountain chain. The band of mixed material just above them is erosional debris dating from that period. The metamorphic rocks were then submerged again, and the horizontal sedimentary beds of sandstone were deposited above them. Finally the entire formation rose above sea level once more and the new soil layers were formed at the top. Hutton cited such examples from his wide-ranging field trips as evidence both of the earth's antiquity and its dynamic activity. In modern terminology a rock formation of this type is called an angular unconformity.

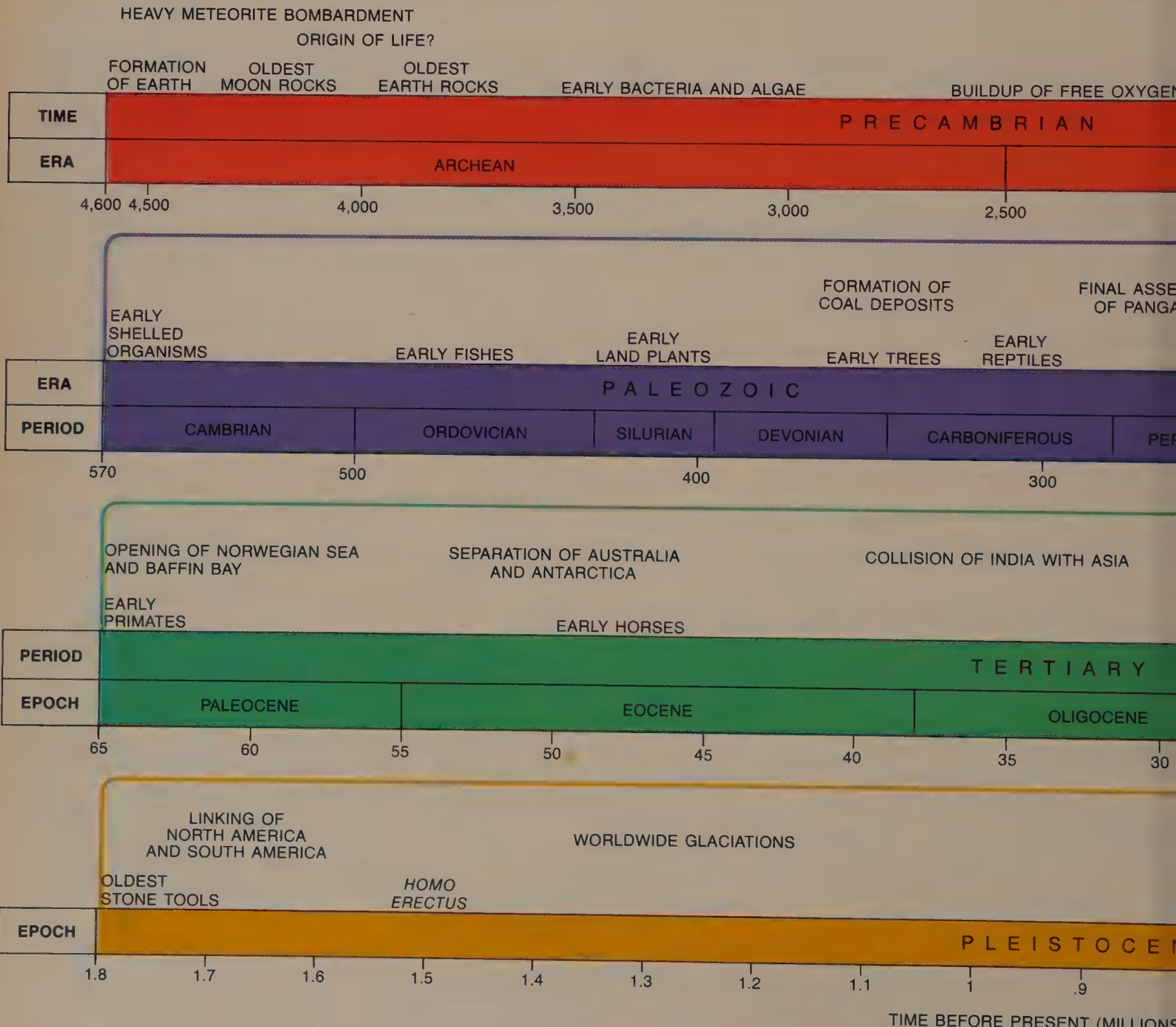


scale folds and faults what an older, obliterated mountain chain was like and how it might have been created by ancient plate motions.

Scales of time involve additional subtleties. A river flowing at several centimeters per second moves at roughly the same speed as a fault block slipping in an earthquake. Yet the distribution of these two movements over a longer period is quite different. The river runs more or less steadily for months or years, whereas the movements along the fault are episodic, with intervals of little or no movement extending over as much as hundreds of years. The distinction between steady and episodic change is central to the current debate over the rate of biological evolution: the key question here is whether or not the geologic time scale embodied in fossil-bearing rocks is precise enough to resolve the issue between the "gradualist" and the "punctuated equilibrium" models of the evolution of species.

Generations of geologists have relied primarily on a biological clock: the sequences of fossils that by their evolutionary changes mark the major divisions of geologic history. In the 19th century that was the only choice; in the 20th century, however, it became possible to calibrate the fossil clock by another timekeeper: the radioactive-decay clock, which depends on the known decay rates of radioactive isotopes of carbon, uranium, potassium, rubidium and neodymium. The scales of the events the two clocks date are apt to be different, as is the nature of the events timed. Working with both clocks is a little like arranging to meet someone at a certain time when one person relies on his pulse to count seconds and the other has a watch with only an hour hand.

Geologists refer to the apparent motion of the sun (or what might be called everyday time) to time comparatively fast processes such as the weather, floods, landslides, volcanic eruptions and earthquakes. They invoke radioactivity alone to time very slow processes such as the evolution of the atmosphere. In between lies the scale of geologic time over which continents move, mountain chains rise, the earth's magnetic field reverses, fossil species evolve

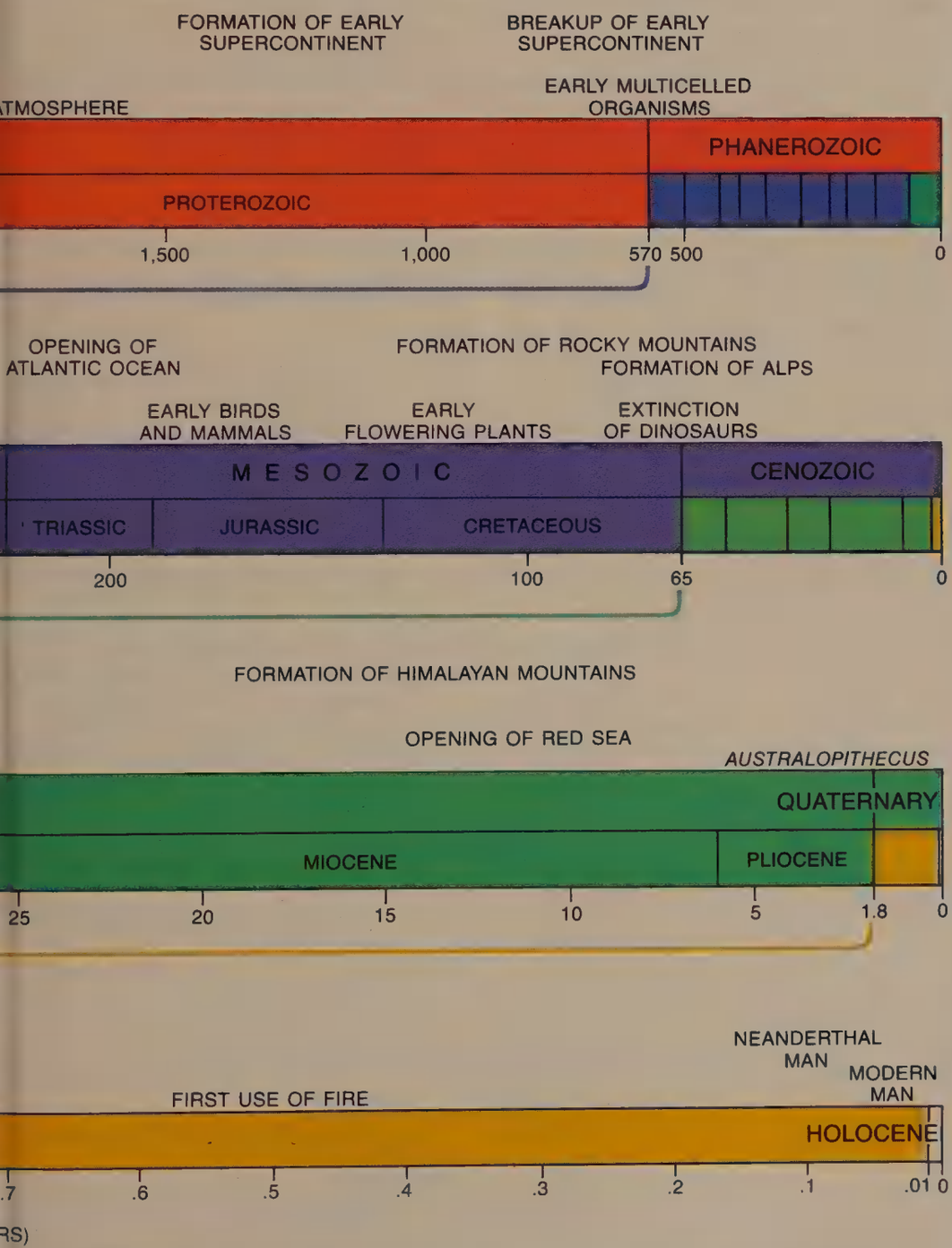


GEOLOGIC TIME SCALE, originally constructed by 19th-century naturalists solely on the basis of fossil evidence, has been calibrated by means of modern radioactive-dating techniques. In this representation the top line shows the full sweep of geologic time from the origin of the earth some 4.6 billion years ago to the present. The comparatively short span of Phanerozoic time, during which fossils of shelled organisms have been abundant in the geologic record, is enlarged in the second line from the top, and successively shorter time segments are enlarged in the next two lines. The three eras of Phanerozoic time (Paleozoic, Mesozoic and Cenozoic) are divided in turn

and glacial epochs last. Over this intermediate time scale reference is made primarily to the specific rock sequences that constitute the geologic record.

For example, in order to study the meandering of a river one can rely on historical records if they go back far enough, perhaps supplementing this information with data drawn from the relicts of prehistoric terraces. To study the long-term evolution of the river, from the early cutting of its valley into bedrock to its later broadening into a characteristic floodplain, however, one has no alternative but to consult the geologic record. The rise in sea level since the retreat of the glaciers some 10,000 years ago is also revealed in the geologic record, which yields information both on the rate of change of the ice sheets and their relation to the sea and on an important property of the earth's interior. As the ice retreated it unloaded the crust, and from the crust's "rebound" it is possible to estimate the viscosity of the mantle material that flowed under it in order to accommodate the rise.

The history of the earth is studied not only for its own sake but also for economic reasons: to explore for oil, gas, metallic ores and other useful substances formed at a specific time and place in the distant past. Simple curiosity, however, is what drives most geologists to want to know just what happened when. They want to know about the last ice age both for what it can reveal about the next one and for a glimpse of what the conditions of life were for early man. Finally, there are those geologists who specialize in the search for evidence that the history of the earth is not just one random thing after another but rather the playing out on a long time scale of the cycles of a great machine.



into 11 periods; the Tertiary period is divided into five epochs, and the Quaternary period consists of the Pleistocene and Holocene epochs. The calibration of the geologic clock by radioactive dating is a continuing concern. For example, according to a recent review of the radiometric evidence by an international group headed by the French geochronologist G. S. Odin, the beginning of the Cambrian period may have been between 540 and 520 million years ago.

If earth history is cyclic, over what periods of time does it repeat itself, and how does one keep track of the process? The hydrologic cycle serves as a model of the cyclic flow of material between different parts of the earth. Water from one major reservoir, the atmosphere, falls on the land and the sea. Some of the precipitation is stored temporarily in the ground and in lakes. The rest follows various routes back to the other major reservoir, the sea. Evaporation from the sea and the land back to the atmospheric reservoir completes the cycle.

The hydrologic cycle is global. Hence by adding up all the water in all the world's reservoirs and flow paths one can arrive at an estimate of the total water content of the system and its major parts. Bypassing analysis of individual reservoirs has the effect of removing heterogeneity: one finds that the global balance is quite steady from year to year. In other words, there is always about the same amount of water in the atmosphere, in the oceans, in the polar ice caps and on the continents. Over periods of less than a year the system may not be so steady, and over periods of a few years the global balance may shift somewhat from one reservoir to another. Indeed, perturbations of the system can elucidate the operation of the cycle. The ice ages represented one such shift in the cycle: water was removed from the ocean and stored in the ice reservoir, and the flows from one reservoir to another were adjusted accordingly; the result was a drastic change in the climate and a lowering of the mean sea level, which exposed much of the area of the continental shelves. The continuing discussion of this major perturbation centers on the possible reasons for the shift and how rapidly the glacial ice caps expanded and then shrank as the earth returned to the comparatively unglaciated stage that prevails today. Polar ice and mountain glaciers still exist, of course, and so it is not possible to know from firsthand evidence what the hydrologic cycle was like when the earth was completely free of ice.

The movement of carbon dioxide through the atmosphere, the ocean and the solid earth provides another opportunity for following the large-scale flow of chemical elements from one part of the earth to another. The carbon dioxide of the atmosphere is taken up by plants in photosynthesis and by rocks in weathering. Photosynthesis manufac-

tures the earth's supply of organic carbon and weathering accounts for the calcium carbonate of limestones, a major product of the transformation of igneous rocks into sediments. Dead plant and animal matter, together with the calcium carbonate shells of mollusks and other organisms, immobilizes the carbon. As these materials are buried in the form of sediment and incorporated into the crust of the earth, carbon is removed from the surface reservoir. At the same time ancient organic matter and limestones are eroded and chemically weathered. By the oxidation of organic matter and the dissolution of calcium carbonate, carbon dioxide is returned to the dynamic system, keeping it in balance.

In much the same way one can construct cycles of all the elements and their isotopes. For example, from the rate at which dissolved calcium enters the ocean from rivers (about 10^{13} moles per year) and the total amount of calcium in the oceans (about 10^{19} moles) one can estimate how long an ion of calcium can be expected to remain in the ocean: about a million years. On the average that is how long it takes before the calcium combines with a carbonate group and settles out of the reservoir as a part of some limestone. The calcium ion may then be buried, perhaps ultimately to be incorporated into a silicate-bearing metamorphic rock. Alternatively, it may go deeper still, becoming part of a magma that will return to the surface in a congealed form as an igneous rock, there to be weathered, dissolved and sent back to the ocean by a river.

These chemical cycles are different versions of the grand geologic cycle that James Hutton, the founder of modern geology, enunciated almost 200 years ago. In Hutton's original version rocks are weathered to form sediment, which is then buried. After deep burial the rocks undergo metamorphosis and/or melting. Later they are deformed and uplifted into mountain chains, only to be weathered again and recycled. In spite of many arguments and theories about the mechanisms of Hutton's rock cycle its basic outlines remain intact as the geologist's way of viewing an earth of constant change.

The continental crust, the repository of the geologic record of the past 3.8 billion years, is engaged in its own cycle of destruction and renewal. Each year some 10^{16} grams of solid and dissolved products of the erosion of the land surface are removed by rivers, wind and glacial ice. Much of the solid detritus is deposited on the continental shelves, but a good deal is lost to the ocean basins. The return is by way of subduction zones, where some of the sediments of the deep-sea basins are scraped off as the oceanic lithosphere plunges into the mantle [see "The Oceanic Crust," by Jean Francheteau, page 114]. The oceanic sediment, which is found in front of oceanic island arcs and along continental margins where the subduction zone borders a continental mass, is plastered back onto the continent. The igneous rocks generated at subduction zones are also added to the continents. Thus over geologic time the continental masses remain in a steady state, even though there are frequent rises and falls in sea level that periodically flood and bare the continental shelves and the lower parts of the continents proper. The sediments and igneous rocks that are added to the continents are welded to them mainly as mountain chains associated with plate boundaries [see "The Continental Crust," by B. Clark Burchfiel, page 130]. Old mountains, then, are the remnants of a past recycling of the continents brought about by the motions of the lithospheric plates.

The recycling story extends deep into the mantle, as crustal materials are thrust to depths of hundreds of kilometers at the subduction zones where plates converge. There they mix with some materials that have never been part of the crust and others that have reached the surface before in their history. Geologists are just beginning to understand how the materials of the earth mix under the conditions of high temperature and pressure prevailing in the interior. The chief tracers of this mixing are the radioactive isotopes of rubidium and neodymium, which are helping to reveal the relations between age and mixing [see "The Earth's Mantle," by D. P. McKenzie, page 66]. Studies of the behavior of the core and the mantle over various time scales will determine to what extent they too can be viewed as parts of the great recycling machine.

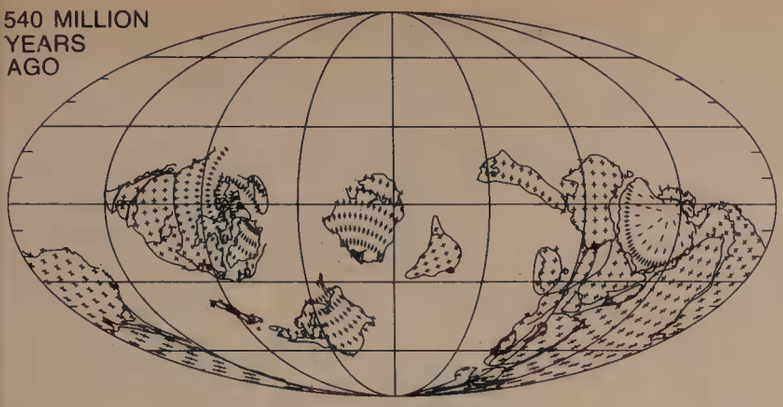
If the machine has been going for billions of years, including some violent perturbations at times, how was it built in the first place? By what series of steady states did it evolve to its present condition? These are the questions that link a fragmentary rock record of the early history of the earth and the other planets to deductions drawn from the astronomy of the formation of the stars and the evolution of the solar system.

Theories of the formation of the solar system out of a nebula of gas and dust are still being refined, but they all have in common the central idea that about 4.6 billion years ago the earth grew to roughly its present size by a combination of two processes: the condensation of primitive solar-nebula material and the accretion of bits and pieces of other planetary matter in the vicinity. The early history of the earth was marked by continued accretion and by a rapid increase in temperature, due to a combination of three effects: heating from the radioactive elements abundant in the early condensate, heating from the impact of infalling material and heating from the contraction of the young planetary body. The rise in temperature, according to the now conventional view, led to widespread melting and a massive differentiation of the earth into a core, a mantle and a crust. All these ideas, proposed long before the exploration of the solar system by spacecraft, were refined in recent years by studies of the moon and of other planets. This was particularly true of the moon, where astronauts were able to sample a body whose history was frozen at an early stage. The moon, with no atmosphere or oceans, suffered no chemical weathering that could obliterate earlier generations of rock. It also did not offer a hospitable environment for life to begin. Viewing it emphasizes how vital the fluid envelope of water and gas was to the earth's machine. The composition of the gas, however, was not the same as that of the atmosphere of today. In its earliest stages the primitive atmosphere was free of oxygen and contained reduced gases such as methane and ammonia [see "The Atmosphere," by Andrew P. Ingersoll, page 162].

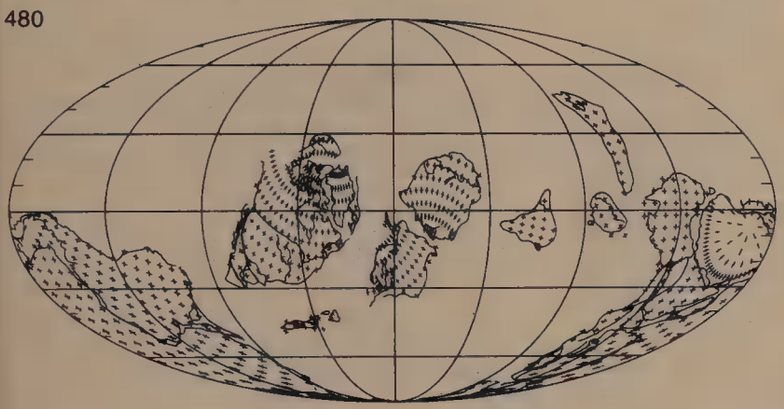
How the dynamics of the earth machine worked in the early days is a deductive study, because no mountains or sediments are preserved to give any hint of the products. The radioactively dated rock record begins in a fragmentary way about 3.8 billion years ago, which is the age of metamorphism by pressure and temperature of a series of originally sedimentary rocks in southwestern Greenland. The record clearly demonstrates that the essential operations of the geologic machine were similar to those in effect today. Iron formations and other types of sedimentary rocks such as sandstones and shales can be recognized in their metamorphosed forms in the Greenland rocks. Igneous rocks found at the site seem to have been created by much the same kind of melting that can be seen operating today. The deformation of rocks is also similar to the deformations of later times.

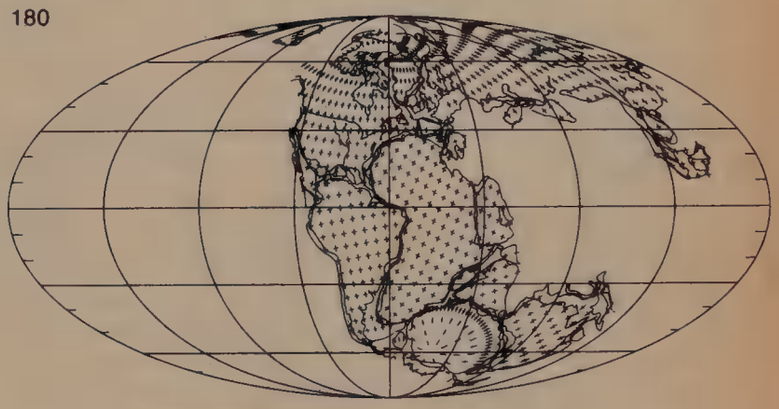
Nevertheless, there are differences be-

CHANGING FACE OF THE EARTH is portrayed at 60-million-year intervals from the Cambrian period to the present in the 10 computer-generated maps on the opposite page. The sequence documents the assembly and breakup of the supercontinent of Pangaea, the two chief episodes of continental drift leading up to the present arrangement of the earth's land masses. The maps were assembled by Alfred M. Ziegler and Christopher S. Scotese of the University of Chicago's Paleographic Atlas Project mainly on the basis of paleomagnetic evidence.


540 MILLION
YEARS
AGO

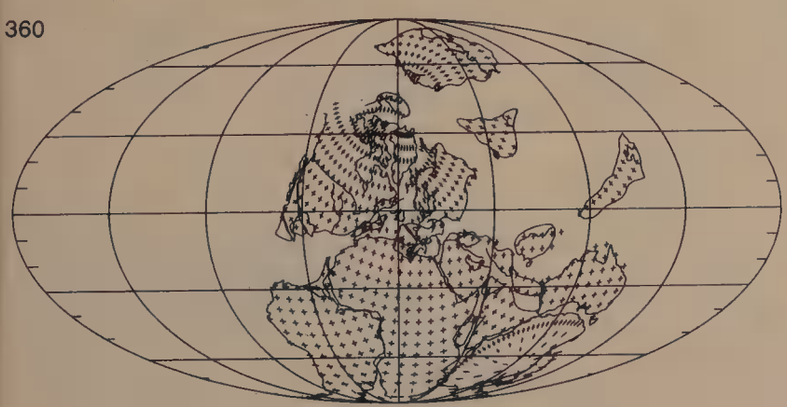

240


480

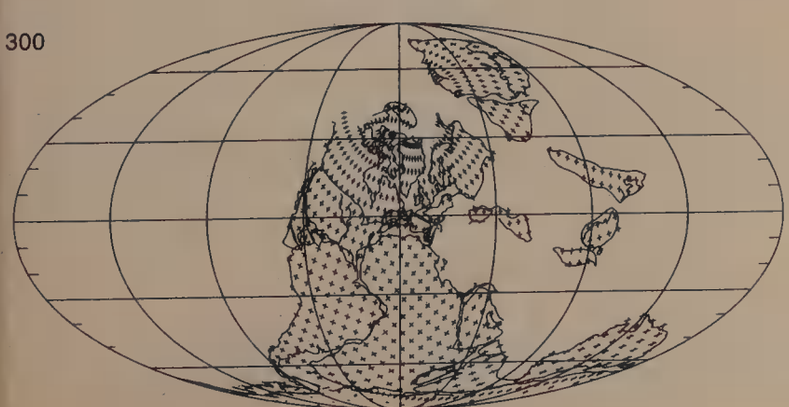

180


420


120


360


60


300

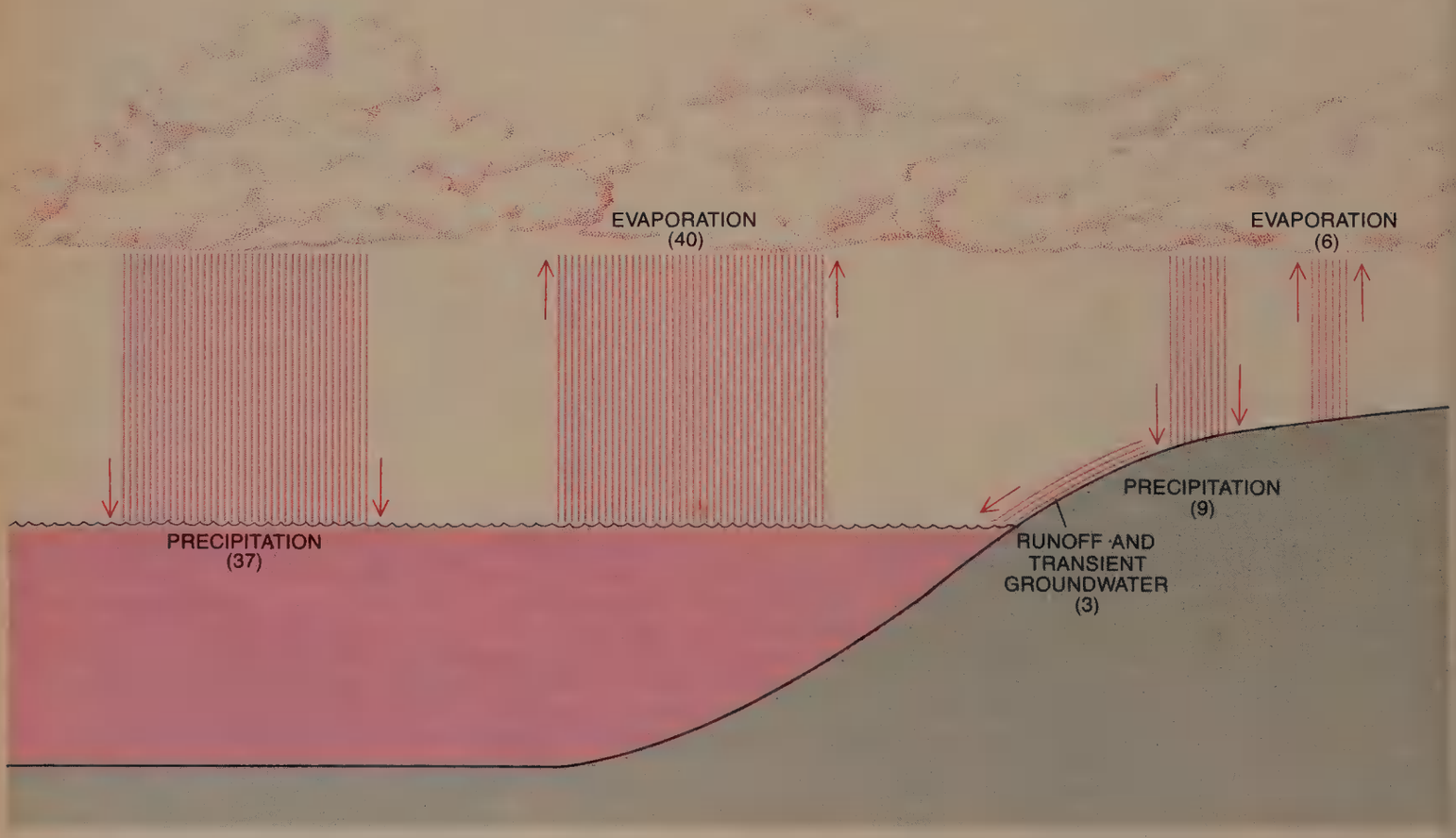

PRESENT

tween these early rocks and more recent ones. Although no fossils have been found in the Greenland rocks (perhaps because the rocks have been so altered by metamorphism), evidence of primitive life is found in somewhat younger ones [see "The Biosphere," by Preston Cloud, page 176]. Some of the basalts found in the earliest earth rocks have compositions that reflect much higher melting temperatures, as if the rate at which the temperature increased with depth in the earth were much greater then than it is now; this is not surprising in view of the earth's early thermal history. Moreover, before about 2.5 billion years ago there were few large masses of granitic rocks and the kinds of sediments formed on shallow continental shelves.

The differences indicate a machine whose internal temperatures were higher, whose atmosphere and oceans were deficient in oxygen and whose land masses were isolated, smaller areas rather than large continents. Yet at all times the average compositions of the rocks and their similarities from one place and time to another demonstrate that the recycling machine was operating at something like the same rate and with the same general character as it does today. It undoubtedly underwent gradual changes, and there are also records of abrupt episodic shifts. The first glaciation of the earth was recorded early in Precambrian time. A more significant change in the cycle came about 2.5 billion years ago, when there seems to have been a sudden rise in the production of granite and the appearance of large continental shelves.

Later in the Precambrian, about a billion years ago, the machine began to look much as it does today. Oxygen in the atmosphere started increasing as photosynthetic organisms grew in numbers and efficiency, and as the march toward more complex life forms accelerated. Although the land surfaces were populated only by algal, fungal and bacterial species, rock weathering in Precambrian soils and the formation of river and lake deposits were proceeding much as they are today; their rates, however, may have been lower. The interior of the earth had by now settled down to something like its present state and through partial melting some regions of the mantle were being differentiated and depleted in certain elements with respect to the mantle as a whole. Whether plate tectonics was the primary mode of heat loss at that time is an open question. The magnetism of ancient rocks gives evidence of continental drift and reversals in the polarity of the earth's magnetic field, and so the dynamo in the earth's core that generates the field must have been in steady operation [see "The Earth's Core," by Raymond Jeanloz, page 56]. Perhaps the plates were too thin to act quite as they do today, or the average size of the plates was smaller. Geologic knowledge of the Precambrian is too fragmentary to give firm answers to these questions yet.

Most geologists working in the past two centuries would have agreed that the biggest change of all was the one at the boundary of Precambrian time and the Cambrian period, marking the beginning of the Phanerozoic eon: the "known" part of the geologic record. It was then that organisms with shells evolved; their fossils make it possible to date the rocks more precisely and to erect the stratigraphic time scale. Compared with the Precambrian there are many more areas of these younger rocks exposed in an unmetamorphosed state, and so one can more readily deduce the course of earth history since then. This change, which is generally believed to have taken place about 570 million years ago, was far more important for the life of the planet than it was for the working of most other parts of the machine, which by then had reached maturity. The parts most affected by the evolutionary development of higher or-



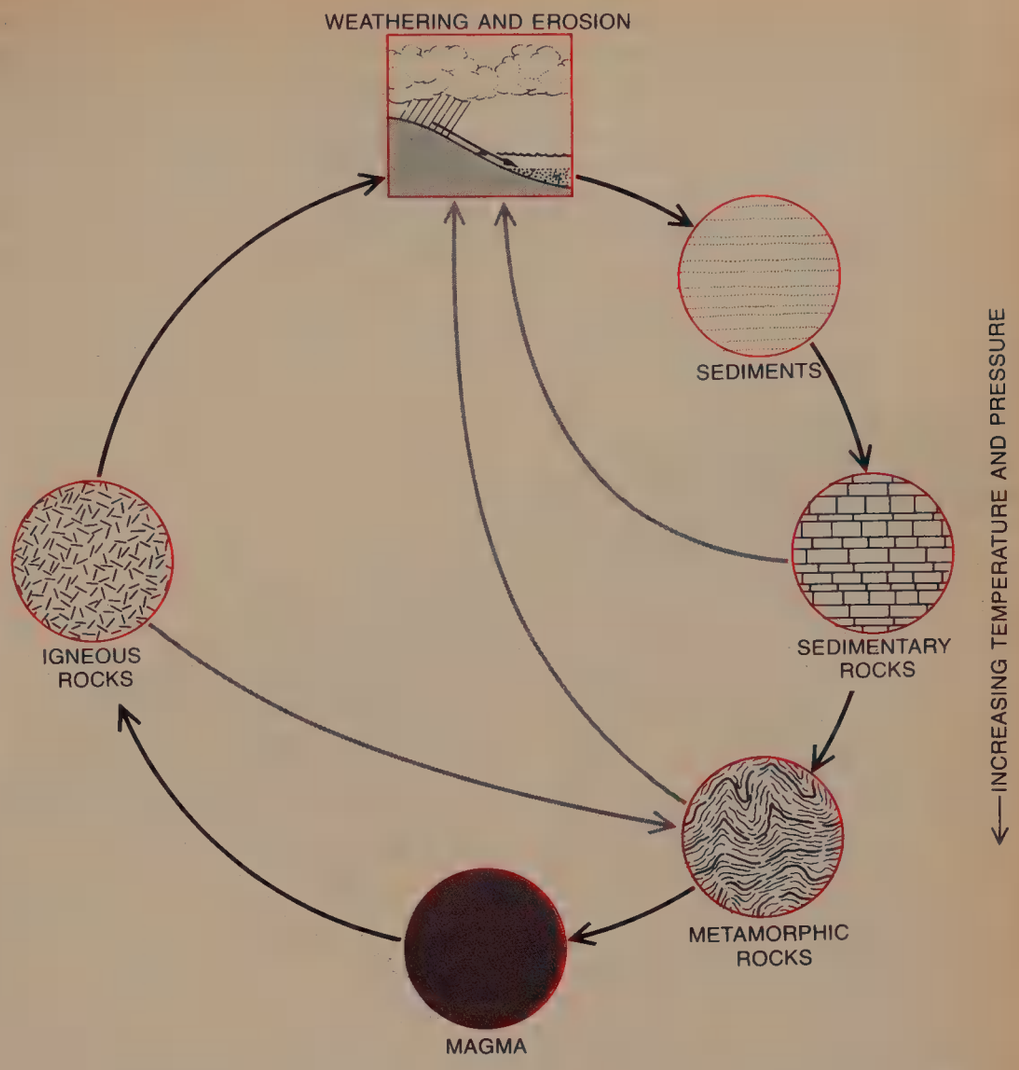
HYDROLOGIC CYCLE, represented quantitatively in this diagram, serves as a model of the cyclic flow of material between different parts of the earth. The numbers shown are multiples of a basic unit equal to 10,000 cubic kilometers of water per year. In spite of occasional short- and long-term perturbations, the global balance among the world's major water reservoirs—atmosphere, oceans, continents and polar ice caps—is quite steady from year to year. In addition to the amounts indicated here water is also brought to the surface of the earth by volcanism. This comparatively small increment consists of former surface water recycled through the interior by way of subduction zones as well as "juvenile" water, namely part of the original water of the mantle that has never escaped to the surface before.

ganisms were those at the surface, where the chemical influence of life processes is felt. The atmosphere arrived at something like its present oxygen level; the deposits of limestone became dominated by the shells of shallow marine organisms, and the chemical cycles of carbon dioxide, oxygen, phosphorus and nitrogen rather abruptly shifted to their present states.

Since that time changes in the rates and quantities of sediments controlled by organisms have been determined by the course of evolution. The vascular land plants first appeared about 400 million years ago and coal beds began to be widespread. The evolution of the flowering plants came about 120 million years ago. The siliceous algae (diatoms) and the pelagic foraminifera evolved about 130 million years ago and started to form the abundant siliceous and calcium carbonate oozes of the sea floor. Much biological sedimentation was shifted to the open sea floor, whose evolution can be understood only for the past 150 million years. The older record of sea-floor spreading has been swept into subduction zones and destroyed by metamorphism or melting.

The earth in old age is predictable from the present running of the machine. As the original supply of radioactive elements steadily decreases, the internal heat of the earth will diminish and simple heat conduction will slowly take over from the currently dominant mode of heat transfer by convection. As the earth cools, the rigid plates of the outer crust will thicken, and eventually they will congeal and become motionless. The "hot spots" that form volcanic centers will cool and solidify. With no internal forces to throw up mountains and move continents around, the external machine on the surface will take over, reducing much of the surface elevations of the land to plains elevated slightly above sea level. Sediment, the debris of erosion, will spread over land surfaces and sea bottom with no interruption by deep-seated rocks brought to the surface. A new balance of chemical elements based on a static tectonic system will lead to a shift in the composition of the atmosphere and the ocean, finally leading, as all mountain building and sedimentation ends, to a new steady state determined almost entirely by life's balance between photosynthesis and respiration.

That balance itself, together with the total mass of biological material, might change significantly as the nutrient reservoirs of the ocean and the atmosphere interact only with the thin surface skin of the earth. This brief essay at describing the future workings of the earth machine is enough to show how dependent the present workings of the planet are on interactions with the interior. No one is sure, however, of how such perturbations in the operation of the present machine can be predicted quantitatively, as is suggested by the current experience with carbon dioxide.



ROCK CYCLE, first proposed by Hutton almost 200 years ago, is still the basis of the geologist's view of the changing earth. In Hutton's version rocks are weathered to form sediment, which is then buried. After deep burial the rocks undergo metamorphosis and/or melting. Later they are deformed and uplifted into mountain chains, only to be weathered again and recycled. The modern theory of plate tectonics is in a sense an elaboration of Hutton's rock cycle.

It has now been almost 100 years since Svante A. Arrhenius, the great Swedish chemist, called the attention of earth scientists to the effects of carbon dioxide on climate and its possible relation to glaciation. For much of that period some workers have been concerned about the steadily rising amounts of carbon dioxide in the atmosphere contributed by the burning of fossil fuels. As the pace of carbon dioxide emissions has increased dramatically in the latter half of the 20th century and as knowledge of the carbon cycle has grown, the issue has come to the point where government commissions and other national bodies estimate how much climatic change there is likely to be and what its effects might be. The comparatively small changes man has worked in the earth machine might have enormous consequences. Large as the earth is, the behavior of one of its biological species may be enough to seriously disturb its balance.

In order to estimate the effects of increased carbon dioxide in the atmosphere one must consider all the fluxes and reservoirs. The human disturbance has been to change a single flux by increasing through coal mining and oil production the rate of return of buried organic carbon to the surface by orders of magnitude over the estimated preindustrial flux. To forecast the consequences one must follow through the entire system the changes that result. More carbon dioxide in the atmosphere leads to a slight increase in global temperatures as a result of the "greenhouse effect." An additional dissolution of calcium carbonate in the ocean and imperceptible changes in the balance of weathering and sedimentation of silicate and carbonate minerals will follow. The warming of the atmosphere will lead to some melting of the polar and glacial ice and a rise in the sea level, and very probably a shift in climatic belts. Reciting this list is only the beginning of a consideration of the innumerable ramifications of a shift in only one flux in

the system. It is this complexity that makes the estimation of the ultimate effects so difficult.

Burning fossil fuels is only the most recent perturbation of the earth machine. Geologic history is in one sense a recital of the many small shifts of the balance, some local and some global, that have characterized this otherwise smooth-running mechanism. The closer one looks at any complex machine, the more clearly one can detect fluctuations in its operation. In the midst of thousands of fluctuations that generate background noise—small shifts in sea level, plate-spreading rates and erosion rates, among others—one can see the signal of large and unusual events, infrequent and sometimes catastrophic. Near the end of the Miocene epoch, about 11 million years ago, the convergence of the African and European plates closed the mouth of the Mediterranean Sea. The Mediterranean dried up and thick layers of evaporites, principally salt and gypsum, were laid down. Soon afterward continuing plate motions reopened the Mediterranean to the Atlantic and it refilled. An immense waterfall existed for a time as water from the Atlantic spilled in to fill up the entire Mediterranean basin. Spectacular as the event was for this area of the world, it probably had a negligible effect on the global cycle; the evaporated water and salt removed from the ocean was minor compared with the huge volume of water in the world ocean.

Salt deposition at another time in the history of the earth was a part of a much more pervasive global cycle, one familiar to all geologists for its relation to massive extinctions of many biological species and almost complete withdrawal of the seas from every part of the continents. The end of the Paleozoic era, the boundary between the Permian and the Triassic periods, about 225 million years ago was marked by the complete assembly of the supercontinent Pangaea with the superocean Panthalassa surrounding it. Over a period of about 200 million years plate motions had gradually joined various pieces of continents into a single land mass. In the process evaporation from the narrow gulfs and bays of closing oceans and from arid regions led to great amounts of salt being withdrawn from the oceans, decreasing the salt content of the open ocean slightly and perhaps altering its density-driven circulation.

As the continents joined, their total perimeter shrank. As a result of mountain building caused by continental collisions in the course of the assembly the continents were dominated by high mountains. The ocean basins were enlarged by the slowing of plate-spreading rates and the lowering of midocean ridges as plate motions became largely limited to the oceanic crust. The consequence of all these effects was a severe restriction of the area of the continental shelves. The diminished habitats of the shallow marine populations of the world and more variability in the climatic regimes of the land, which had recently gone through another glacial age, set the stage for the rapid evolution of new species and the extinction of old ones. A lot of guesswork is needed here because the rock record for the period is sparse. Most of the sediments deposited on the continents have since been eroded and the sea-floor record has been swept into the subduction zones.

Even this unusual time was soon superseded by a return to normal conditions as the supercontinent proved unstable. It began to rift into separate continental masses, and the Atlantic and Indian oceans opened. The seas transgressed the continents, salt beds were eroded, the salt returned to the ocean and the normal pattern of life resumed on the continental shelves. The episode demonstrates the machine's ability to right itself and settle down to a smooth operation after a violent shudder.

The glacial ages represent another kind of perturbation, one that primarily affected the surface and near-surface layers of the earth. The earth has been glaciated several times in its history, at least twice during the Precambrian, once in the early Paleozoic, once near the end of the Paleozoic and many times in the Pleistocene, which ended only about 10,000 years ago. Each of these events may have been triggered by continental-drift patterns that put a continent in a polar region and caused shifts in oceanic circulation. Perhaps there was a correlation with a wider than normal fluctuation of the carbon dioxide level. In each glacial period the periodicity of changes in the earth's orbit around the sun led to oscillations in the advance and retreat of the ice; the oscillations, called Milankovitch cycles, are now thought to account for the glacial and interglacial stages of the most recent ice ages. Each time, the postglacial earth returned to its preglacial state with little change, leaving behind a record of glacial sediment and grooved and striated bedrock, the pavement over which the ice moved.

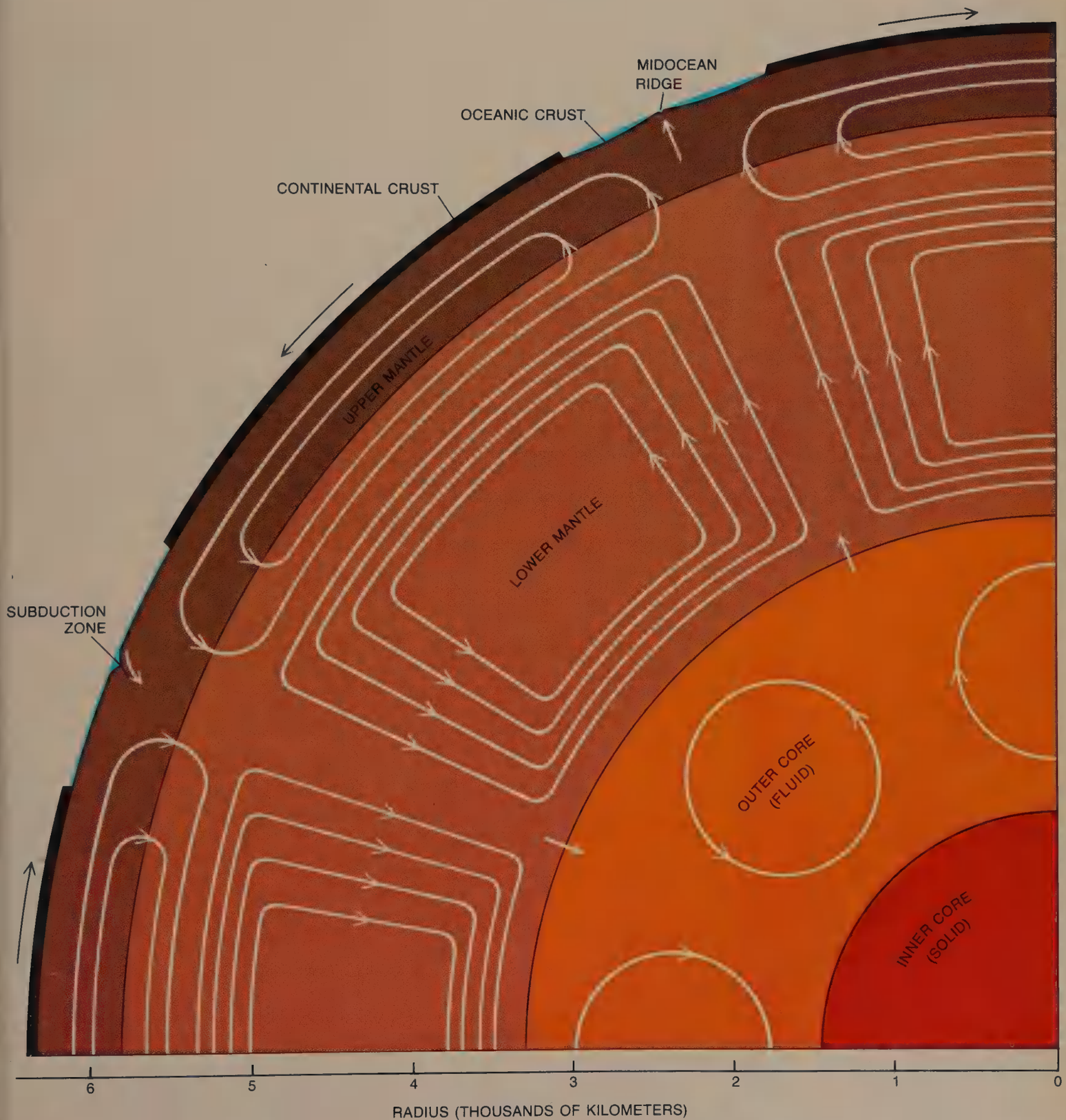
In the past few years geologists have begun to evaluate disturbances of the earth's balance due to extraterrestrial sources. A thin bed of iridium-enriched clay has been found in a number of places around the world where sediments were deposited during the transition from the Cretaceous period to the Tertiary, about 65 million years ago. It has been proposed by Luis W. Alvarez, Walter S. Alvarez, Frank Asaro and Helen V. Michel of the University of California at Berkeley that this bed is the product of the impact of a large meteoritic body. Among the postulated consequences of the collision (with a body estimated to weigh about 10^{18} grams) were clouds of dust that obscured the sun, sudden cooling of the atmosphere and the ocean, changes in the chemistry of the atmosphere and the deposition of a thin layer of clay enriched in certain rare metals such as iridium. These changes are thought to have led to the extinction of many plant and animal species in the sea and the dinosaurs on the land.

Regardless of the durability of this hypothesis as an explanation of the geochemical and biological associations of the boundary between the Cretaceous and the Tertiary, geologists have come to recognize that collisions with asteroid-size bodies have been an important feature of the earth's history since its formation 4.6 billion years ago. Such heavy impacts may have come every 100 million years or so. Most of the events have left little trace because of erosion over the next few tens of millions of years, but some have left conspicuous circular scars.

As the interconnectedness of the huge and complex earth system becomes more widely recognized, earth scientists are learning how to link up subsystems and explore new relations. Oceanographers have combined marine chemistry, marine biology and the geology of the sea floor, from midocean ridges to volcanic-island arcs, to see the oceans as an entire system. Indeed, it is the only way to estimate the interactions of the oceans with the carbon cycle and the carbon dioxide greenhouse [see "The Ocean," by Wallace S. Broecker, page 146]. With geologists and paleontologists they are now mapping the past of the oceans, particularly in relation to the recent glacial epoch. Predicting the operation of the surface part of the machine, powered by solar radiation, has become the business of oceanographers, geologists, atmospheric scientists and biologists. The ability to predict is dependent on devising ever more inclusive and sensitive models of the system and testing them against observations. Geophysicists working on the dynamics of the interior heat engine are engaged in correlating its structure and constitution, as revealed by earthquake waves, with its mineralogical and chemical constitution, as deduced from laboratory experiments on rock systems at high temperatures and pressures. Isotope-ratio studies of rocks derived from the interior and brought to the surface provide observations that tell about mixing mechanisms in the mantle. Theoretical studies of convection dynamics in plastically deformable solids at high pressures are showing how the mixing proceeds. The core, the mantle and the crust are no longer regarded as separate do-

mains but as interacting parts of a larger system whose properties and dynamics are modeled by geophysicists, geochemists and petrologists.

Earth scientists have long been familiar with the ways in which the actions of the interior affect the exterior through volcanism, mountain building, the flow of heat and the geomagnetic field. Now they are seeing how chemical weathering and the differentiation of the materials of the interior as they are brought to the surface react on the interior as the altered material is plowed back into the mantle by subduction. In this way the surface machine is further coupled to the interior machine. As the subsystems are linked the earth may come to be thought of more in the ways one thinks of a highly differentiated organism: as a system so complex that the ultimate reduction to simple forces and bulk compositions does not lead to a satisfactory understanding of the wonderful diversity and detail that can be observed directly at the surface and sensed remotely in the interior.



LARGE-SCALE MOTIONS of the major parts of the earth are indicated by arrows in this highly schematic diagram. Heat-driven convection in the fluid outer core has a dynamo effect that is responsible for the geomagnetic field. Convection in the upper mantle drives plate tectonics. Volcanism transports molten material to the surface at midocean ridges and other places. Solid material is returned to the interior at subduction zones. The degree of mixing between the upper mantle and the lower mantle is a subject of debate; in this case a model calling for separate convection cells has been adopted. Mixing of material between the lower mantle and the outer core is still speculative.

The Earth's Core

Indirect evidence indicates that it is an iron alloy, solid toward the center but otherwise liquid. It is the turbulent flow of the liquid that generates the earth's magnetic field

by Raymond Jeanloz

It is ironic that one of the great spectacles of terrestrial nature, the light of the aurora shimmering in the night sky, is a clue to the character of the earth's hidden and enigmatic core. The ultimate cause of the aurora is the interaction of the magnetic field generated in the core and the "wind" of electrically charged particles flowing outward from the sun. The core also has much to tell about the earth's formation and geologic history. Indeed, there are indications that it may still be influencing the distribution of temperature in the overlying mantle and thus may indirectly govern large-scale geologic processes at the surface. It is also clear that the composition of the core is a major factor in any model of the bulk chemistry of the earth.

The present nature of the core is best determined from seismological data, that is, the information gathered by studying the acoustic waves generated by earthquakes. Such data reveal that the core extends from a depth of about 2,900 kilometers (1,800 miles) to the center of the earth at 6,370 kilometers (3,960 miles). They also show that the inner core is solid, with a radius of about 1,200 kilometers, and that the outer core is liquid. What the composition of the solid and the liquid might be are matters to which I shall return.

Conditions are known to be extreme at these depths. The pressure ranges from 1.3 to 3.5 million atmospheres, which is to say from 1.3 to 3.5 million times the atmospheric pressure at the surface of the earth. Temperatures are estimated to be in the range from 4,000 to 5,000 degrees Celsius (7,200 to 9,000 degrees Fahrenheit).

The most direct information on the core in the past comes from studies of the ancient magnetic field of the earth. For example, the magnetization of some of the oldest rocks suggests that whatever process generates the magnetic field in the core was already operating 3.5 billion years ago. Students of such matters are still far from a thorough understanding of the geomagnetic field: how it is generated, when and how it started and how it has evolved. Nevertheless, it is now possible to begin to understand how the core might have formed and evolved and how these processes may have affected the geologic evolution of the earth.

The geomagnetic field is often described as being like that of a dipole. In other words, it looks like the field that would be generated by a bar magnet at the center of the earth, with the lines of force looping from the south magnetic pole to the north. This is actually a parochial description, because the magnetic field is like that of a dipole only near the surface of the earth, where it can be studied most readily.

Around the earth in the magnetosphere the lines of force in the magnetic field are strongly distorted by the solar wind; they are crushed toward the earth on the side facing the sun and swept far into space on the night side of the planet. Similarly, all models of the source of the magnetic field show distorted field lines in the core. Nevertheless, it is worth remembering that about 90 percent of the field now observed at the surface is dipolar. The rest consists of a more complex pattern of field lines that can be described in terms of several poles rather than just the two that describe most of the present field.

Since the pioneering work of Walter M. Elsasser of Johns Hopkins University, Edward C. Bullard of the University of Cambridge and others the geomagnetic field has been understood as originating with magnetohydrodynamic processes in the earth's liquid outer core. In general terms the processes entail convection in an electrically conducting fluid, with the result that the core acts as a dynamo maintaining and regenerating the magnetic field. Specifically, as the field lines directed toward the center of the earth (the poloidal lines) enter the outer core they are pulled in the direction of the earth's rotation. The rotation of the solid inner core probably tends to wrap the field lines around the earth's axis (producing a toroidal component).

Furthermore, it is thought the field lines become contorted by smaller-scale cyclonic motions that result from the fact that the core is rotating along with the rest of the earth. The cyclonic motions are analogous to the hurricane patterns that arise in the atmosphere. Although the contortions of the magnetic field lines play a central role in current theories of the magnetic dynamo, neither the exact origin nor the detailed pattern of the contortions is known. It is worth emphasizing, however, that without the dynamo process the magnetic field would certainly die out within 10,000 years or so. Therefore the field must be continuously maintained or regenerated by the fluid motions.

In order to better understand the nature of the magnetic field within the core it is necessary to know the convective-flow pattern of the liquid. The trouble is that the magnetic field can significantly modify the flow generating the field in

MAGNETISM OF THE EARTH is visible in these images of the aurora made from data transmitted by the spacecraft *Dynamics Explorer 1*. The aurora is the circle at the lower right in each image. It is created when electrically charged particles in the "wind" of the sun's expanding atmosphere are trapped in the earth's magnetic field. Plunging into the earth's atmosphere along the magnetic lines of force descending toward the magnetic poles, the solar-wind particles interact with molecules in the earth's atmosphere to generate the light of the aurora (here actually detected at ultraviolet wavelengths). The bright area at the upper left in each image is the side of the earth lighted by the sun. Each image represents data received for a 12-minute period. The images show a hitherto unknown configuration of the aurora, in which the circular area of emission is crossed by a linear one; the configuration is called the theta aurora because of its resemblance to the Greek letter. The images were obtained with the University of Iowa's auroral-imaging instrumentation. They are provided through the courtesy of Louis A. Frank.



the first place. Hence no one has yet fully solved the problem of determining the fluid motions in the outer core.

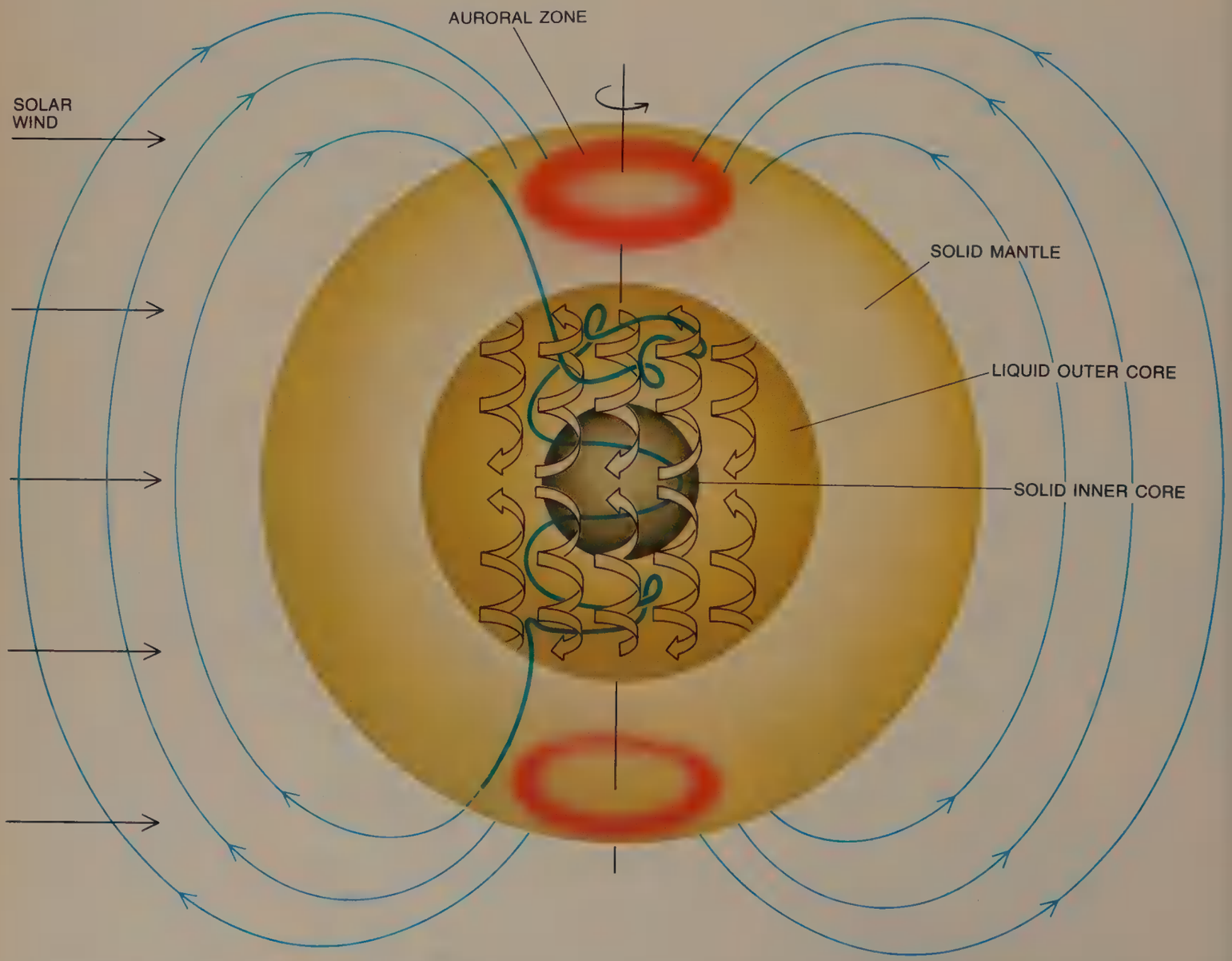
A logical starting point, however, is to first ask what the flow would look like if there were no magnetic field. Possible answers to this purely hydrodynamic question are emerging from the theoretical and experimental work of Friedrich H. Busse and his colleagues at the University of California at Los Angeles. They find that both the presence of the solid inner core and the rotation of the earth are major influences on the flow pattern in the liquid outer core. In conditions thought to be appropriate to the earth's core the pattern takes the form of screwlike rollers. One can speculate that if this pattern of flow is not too strongly modified by the presence of the magnetic field, the rollers are intimately associated with the contortion of the field lines that is required for the dynamo.

The magnetization that is retained in rocks of different ages has been studied by, among others, M. W. McElhinny and his colleagues at the Australian National University. They find that a geomagnetic field of about the same intensity as the present one existed at least 2.5 billion years ago and probably existed 3.5 billion years ago. (The latter determination was made on the basis of only a few samples.) Although data are absent for a substantial fraction of the earth's history, inasmuch as the planet is 4.6 billion years old, it seems that the geodynamo started operating fairly early in geologic time.

The fact that the presence of an inner core appears to be important in the process generating the magnetic field leads to the further inference that this solid region at the center of the earth has existed for at least 3.5 billion years. Even though the inner core may have grown or shrunk since then, there is no reason to infer that the general nature of the flow in the outer core has changed substantially throughout recorded geologic time. In other words, the magnetic evidence suggests that the basic structure of the liquid outer core surrounding the solid inner core has existed for at least three-fourths of the earth's history.

The understanding of the geomagnetic dynamo and of the convective-flow pattern in the core has advanced considerably in the past several years. It is reasonable to predict that the theoretical and experimental work now being done will lead to major new insights. Nevertheless, certain aspects of the problem are still baffling.

A good example is the reversal of the magnetic field that has occurred hundreds if not thousands of times in geologic history. In a reversal the north magnetic pole switches from pointing toward geographic north to pointing south (or vice versa). Reversals appear not only in the earth but also in the sun and even in laboratory dynamos, presumably in response to the chaotic nature of magnetohydrodynamic processes. On a practical level the frequent and apparently random geomagnetic reversals have provided an exceedingly useful clock for timing geologic events and correlating geologic deposits. In-



MAGNETIC FIELD IS GENERATED by a dynamo in the core. The details of how the dynamo works are not known, but in the model depicted here it is assumed that the electrically conducting metallic liquid of the core flows in screwlike rollers. The lines of force in the magnetic field would be threaded through the rollers; here a single such line is depicted. It is the thick line from north to south.

deed, it is the magnetic reversals recorded in the rocks of the ocean floor that have yielded one of the main lines of evidence for the theory of plate tectonics.

It is only fairly recently that high-quality data have become available to trace in detail what happens when the magnetic field reverses. Apparently the magnetic pole follows a convoluted path at the earth's surface. It is clear, however, that the field can reverse completely within just a few thousand years.

Several models have been proposed for what the field should look like during a reversal. According to one of them, as an example, either the north or the south magnetic pole should appear to follow a line of constant longitude, with the pole crossing through any given sampling location on the surface of the earth in the course of a reversal. One can visualize the magnetic pole as splitting into a hoop that sweeps the surface in a north-south direction. The data now available, however, do not support this hypothesis.

A major difficulty in modeling the field as it reverses is that during the process the absolute intensity of the field is reduced to about 10 percent of its normal value [*see illustration on next page*]. At present the geomagnetic field is about 90 percent dipolar, the rest being multipolar. No one is certain what fraction of the reduced field is dipolar while a reversal is under way. To derive the poles shown in the illustration it is necessary to assume that the field was dipolar during the depicted reversal. Perhaps, however, there are effectively several magnetic poles existing simultaneously during a reversal. If they fluctuate in strength, they can give rise to the erratic path seen in the illustration. This possibility can be investigated only by studying the same reversal at several different places.

Reversals seem to be more than just a passive dying out and rebirth of the field because they appear to happen on a time scale that is much shorter than the time it would take for the field to be regenerated. Hence one may ask: Are reversals caused by turbulence or shifts in the detailed flow patterns in the outer core? No one has been able to answer the question. It is quite possible the dynamo is self-reversing, that is, a reversal can be initiated internally, without any external trigger.

Regardless of the mechanism, one consequence of reversals is only now beginning to be recognized. What is being found in the growing new discipline of biomagnetism is that complex organisms synthesize magnetic components whereby their behavior can be significantly affected by changes in the geomagnetic field. Therefore one may speculate that reversals play a role in biological evolution.

The geologic record of the earth's magnetic field yields information on the nature of the core in the past. In order to proceed further in understanding the core it is necessary to consider the available data bearing on its present nature as well. The only detailed and direct information comes from seismological studies, which provide values of the density of the material and the velocity and attenuation of sound as a function of depth in the earth. "Sound" is employed here in a loose sense to refer to the mechanical waves produced by earthquakes (or large manmade explosions) and propagated through and around the globe. Such waves are of low frequency compared with audible sound: from about 10^{-4} to 10 hertz, or cycles per second, which is some 100 to a million times lower in "pitch" than the concert A. From these data the pressure can be calculated at each depth. Moreover, liquid and solid regions can be distinguished because a liquid transmits only compressional waves (waves moving back and forth in the direction of their travel), whereas a solid transmits both compressional waves and shear waves (waves moving at right angles to their direction of travel).

The basic structure of the earth implies a crude but strong constraint on the deep temperature. Evidently the geotherm (the average temperature as a function of depth) is below the melting point of the inner core and the mantle, since they are both solid. By the same token the geotherm is above the melting point of the outer core.

Usually this argument is taken one step further on the assumption that the inner core is solidified outer-core material. The assumption is plausible and is not contradicted by any data now available. No one has yet demonstrated, however, that the inner and outer core are in chemical equilibrium, as the assumption would imply. In fact, the inner core may be in chemical disequilibrium with the outer core, which in turn may be (and is often taken to be) in disequilibrium with the mantle. In other words, the composition of the inner core may not be related in any direct way to the composition of the outer core.

If the inner core has formed (or is still forming) by crystallization from the liquid outer core, the boundary between them is fixed within the temperature interval of melting and solidification of core material at a pressure of 3.25 million atmospheres for the present size of the inner core. For example, the inner core could be pure iron that has crystallized out of the alloy liquid of the outer core. The temperature at the boundary between the inner and the outer core is similarly constrained even if the outer core is heating up and hence is growing by the melting of the solid inner core. In either case one would in general expect a region of partial melting between the inner and the outer core, because only the simplest chemical systems melt at a sharply defined temperature, that is, begin melting or begin solidifying at the same temperature.

The core is thought to be a complex alloy, which would therefore melt over a range of temperatures. As a result there has been much interest in recent seismological studies by V. F. Cormier of the Massachusetts Institute of Technology and G. L. Choy of the U.S. Geological Survey, who find evidence for a "mushy" zone of reduced velocity and relatively high attenuation of seismic waves in the top few hundred kilometers of the inner core. Their conclusions are uncertain because of the great difficulties involved in resolving such details of physical properties near the center of the earth. It does appear, however, that there may be an anomalous zone such as one would expect for a slurry of liquid and crystals. This is circumstantial evidence that the top of the inner core is right at the melting point. Thus if the composition of the core were known, an experimental determination of the melting point of this material at the pressures of the core would yield a direct determination of the temperature near the earth's center.

It is also possible to infer that the temperature has not varied enough to completely melt (or solidify) the core within the past 2.5 to 3.5 billion years. The inference is based on the record of the earth's magnetic field suggesting the inner and the outer core have existed for at least that length of time.

The core is usually represented as consisting mainly of iron. This interpretation is clearly in agreement with the seismological data, but two other lines of evidence can be invoked to strengthen the conclusion. The first is that the generation of the magnetic field requires the core to be metallic (that is, electrically conducting) in order for the geodynamo to operate. The second is that no other element having the observed properties of the core is abundant enough in the cosmos to be a plausible candidate.

Hence there is a gross separation of the earth into an iron-rich region (the core) and a silicate region (the mantle and the crust). Silicates are the complex silicon-oxygen compounds making up rocks, so that the two regions are reminiscent of the two general classes of meteorites: iron and stony. Beyond this, however, there is no evidence for a more detailed analogy between the earth's core and the nature of meteorites. In fact, it is not possible that the core is made up only of iron or the nickel-iron alloy commonly observed in iron meteorites. This is evident from a comparison of the density of iron alloys at high

pressures with the density of the core: a small amount of a component less dense than iron, such as sulfur, oxygen or silicon, must also be present.

There is no consensus on the composition of the core other than that it is predominantly iron. In part the reason is that gross physical properties, such as the measured densities and seismic velocities within the core, cannot be exploited to uniquely determine the core's chemical composition. Moreover, only a small amount of any of the alloying elements that have been proposed is required in order to match the observed properties of the core. (Typical values are about 8 to 10 percent by weight.) Finally, it is quite possible—probable according to many investigators—that the core alloy has many components in addition to iron.



REVERSAL OF THE GEOMAGNETIC FIELD gives clues to the action of the dynamo in the core. A reversal that took place 15 million years ago is traced here on the basis of the magnetism recorded in a sequence of lava flows at Steens Mountain in Oregon. The switch was from reverse to normal, with normal meaning that magnetic and geographic north are in the same direction. The north magnetic pole followed a convoluted path, recorded here for a period of 15,000 years; each filled circle represents a separate measurement of the pole position. The field's strength and direction are indicated at approximately 500-year intervals by the colored arrows. During the reversal the geomagnetic field did not necessarily have two poles, as it does today, so that the poles depicted here represent only a schematic view. The erratic path of the pole could result from variations in the dipole and nondipole components of the field or from shifts in field's main direction. The data on which the drawing of a magnetic reversal is based are from Robert S. Coe of the University of California at Santa Cruz, M. Prévot of the University of Paris and E. A. Mankinen and Charles S. Grommé of the U.S. Geological Survey.

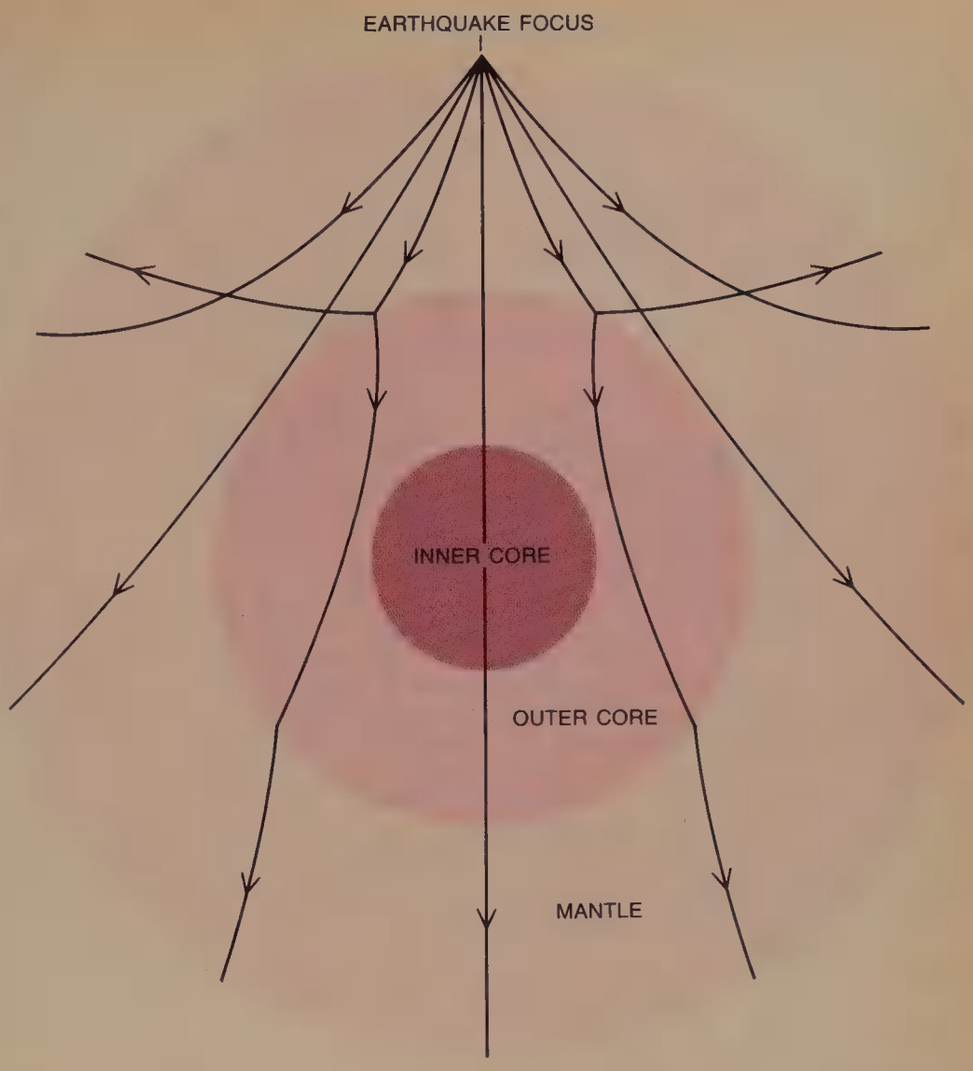
In view of these uncertainties it is best just to summarize contemporary thinking by saying the present favorites for the main light components in the core are sulfur and oxygen. Sulfur was proposed by V. R. Murthy and H. T. Hall of the University of Minnesota primarily because it is depleted in the rest of the earth in relation to cosmic abundances. If enough sulfur could be assigned to the core, the planet as a whole might turn out to be undepleted in sulfur. It is relevant that iron sulfides are found in meteorites.

The trouble with this argument is that sulfur is relatively volatile, and it has long been recognized that the bulk earth is depleted in volatiles in relation to cosmic abundances (most notably in hydrogen and helium but also, for example, in potassium). In addition the high-pressure data indicate that only about 8 percent by weight is required for sulfur to lower the density of iron to the density of the core. That amount of sulfur is not nearly enough to make up the deficiency of the bulk earth.

This relation weakens the original justification for invoking sulfur as the lighter component, but it does not rule sulfur out. Indeed, an iron sulfide combination is regarded by many workers as being the most plausible one for the core. It is worth noting that iron sulfide is a good electrical conductor and that it melts at temperatures several hundred degrees below the melting point of mantle minerals. Its properties are therefore consistent with those of the core, and one can see how the bottom of the mantle could be solid (a silicate with a high melting point) and the outer core liquid (a sulfide with a lower melting point).

The main proponent of oxygen as the light component in the core is A. E. Ringwood of the Australian National University. He suggests that at high pressures iron oxide becomes metallic. The point is crucial because at low pressures iron oxide is not metallic. The required metallization would necessitate a drastic change of properties.

There is as yet no clear evidence by which to judge Ringwood's hypothesis, although shock-wave experiments in the laboratory have demonstrated that at something less than a million atmospheres of pressure iron oxide does undergo a transformation. Unfortunately in the experiments the nature of the transition could not be determined. In the transition does the bonding become metallic, or is there just a change in crystal structure? Answers to this question and others about the properties of iron oxide at high pressures await the conclusion of additional experiments now in progress. In any case the proposed metallization of iron oxide should not be regarded as improbable. After all, the evidence is that even oxygen can become metallic at high pressure, and molten iron oxide appears to be a semimetal at high temperatures.



SEISMIC WAVES provide data on the physical properties of the core. Two types of waves move through the earth from the focus of an earthquake: *P*, or compressional, waves (waves that move back and forth in their direction of travel) and *S*, or shear, waves (waves that move at right angles to their direction of travel). *S* waves, which cannot travel through the body of a liquid, do not pass directly through the core, demonstrating that at least the outer core is liquid. *P* waves go through both solids and liquids. Besides information on the state of the core, seismology yields data on density, which make it possible to calculate the pressure at each depth.

One of the main differences between the oxide and the sulfide hypotheses is that under the oxide hypothesis the core must have acquired its present composition at high pressures. Ringwood concludes that below the metallization pressure oxygen does not combine with iron in any significant amount (in relation to the silicates of the mantle). This is why the core would have to form at high pressures to incorporate oxygen. In contrast, sulfur can readily be alloyed with iron at low pressures.

The effect of combining either sulfur or oxygen with iron is that the melting point of the compound is lowered. At low pressures sulfur has a much larger effect than oxygen on the melting point of iron, and it is thought this difference may persist at high pressures. Therefore a notable difference between the sulfide and oxide models for the core is that melting would begin at significantly lower temperatures in a sulfide composition than it would in an oxide one. As a result it may be easier for a core to start forming if it is sulfur-rich than if it is oxygen-rich.

In order to apply this information it is necessary to know the melting temperatures of iron alloys at the pressures of the core. The melting point of iron itself has recently been determined in the shock-wave experiments of J. M. Brown and R. G. McQueen at the Los Alamos National Laboratory. For the first time they were able to discern the onset of melting in pure iron at 2.5 million atmospheres.

Applying these data, Brown and McQueen have modeled the melting in iron alloys at pressures corresponding to those at the boundary of the inner and the outer core. On the assumption that this boundary corresponds to the melting-freezing transition and that the core is iron sulfide, they are able to estimate temperatures throughout the core. For example, they arrive at a value of 3,700

degrees C. (plus or minus 500 degrees) at the core-mantle boundary, which is close to previous estimates. This is the first time estimates of the temperature near the earth's center have been bracketed by experimental data obtained at the conditions existing within the core. Experiments that are now being done to determine the melting behavior of alloys at high pressures should further clarify the possible range of temperatures in the core.

A subtler connection between composition and temperature at the core is related to the source of energy that powers the geodynamo. Two distinct mechanisms—thermal and compositional—have been proposed for driving the convective flow that generates the earth's magnetic field within the outer core.

The thermally driven flow requires that the fluid be heated by a local source of energy. Warmer regions of the outer core would then rise because they are less dense than the colder regions, which sink. This is the familiar type of convection that occurs in the atmosphere, in a pot of water on the stove and (on a longer time scale) in the earth's mantle.

The compositionally driven flow is different in that dense and less dense regions are formed even if there are no differences in temperature. It is simply an unmixing process of the kind that can take place in a mixture of oil and water; the oil rises and the water sinks. David Gubbins of the University of Cambridge and D. E. Loper of Florida State University have suggested solidification of the outer-core fluid could result in dense crystals that would sink toward the inner core while the remaining, less dense liquid would rise toward the top of the core. This separation process could apparently be quite effective in driving the flow that generates the magnetic field. The details are not well understood, however, and most workers consider this model of the dynamo to be speculative.

The thermally driven dynamo does not call for any compositional difference between the inner and the outer core. All it needs is a source of energy. One possible source is the decay of radioactive isotopes such as uranium 238 or potassium 40, which are present in the mantle and the crust. Are such elements also present in the core? Recent studies of the uranium content of minerals by D. S. Burnett and his colleagues at the California Institute of Technology suggest that sufficient amounts of radioactive uranium are capable of partitioning into core-forming metal to be an important source of heat. Furthermore, both theoretical results and new experimental data support the idea that the chemical bonds of potassium change at high pressures in such a way that radioactive potassium could also combine with the metallic iron of the core.

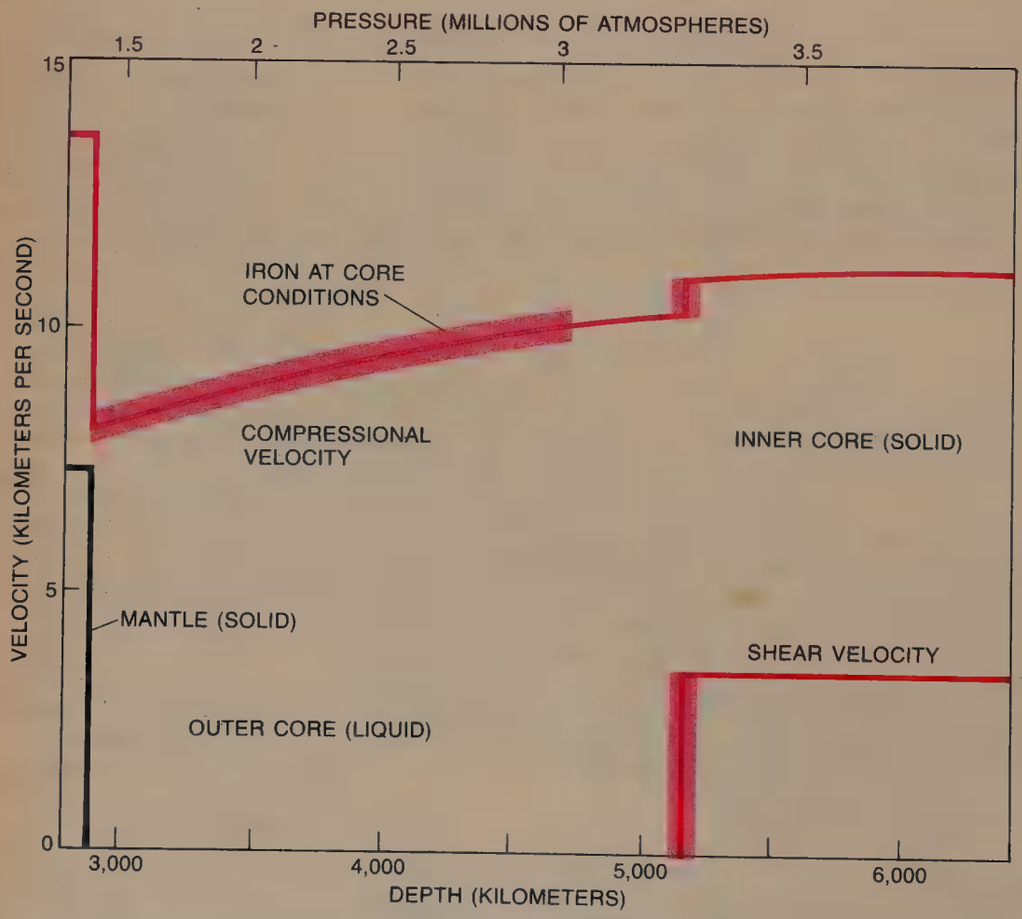
In each case the data are too scanty to allow anything more than speculation about heat sources in the core. The point is that according to present knowledge radioactive decay could be the dominant source of energy driving the flow in the core. Incidentally, one reason some geophysicists argue for significant amounts of potassium in the core is that it could partly explain why the mantle and crust are depleted in potassium in relation to cosmic abundances.

There may be other sources of energy to drive convection in the outer core. For example, considerable heat could have been released when the earth originally assembled or when the core formed. This "primordial heat" hypothesis, to which I shall return, is intimately associated with whatever view one might hold on how the earth was formed.

Another possibility is that if the inner core is freezing out of the surrounding liquid, there could be enough heat from the latent heat of crystallization to power the geodynamo. This hypothesis has been discussed extensively by John Verhoogen of the University of California at Berkeley, who emphasizes the uncertainties in such a model arising from the lack of data on melting in complex alloy systems at high pressures.

Whatever the precise source of energy may be, the fundamental instability driving thermal convection is that less dense fluid lies under denser fluid. This situation arises because there is a sufficient increase, on the average, of temperature with depth and because thermal expansion causes the density of materials to decrease as the temperature is increased by the heat sources.

If the geodynamo is thermally driven, the temperature of the vigorously flowing region in the outer core increases adiabatically with depth (that is, with pressure). In an adiabatic process the energy content of a given parcel of fluid remains the same, which is to say there is not enough time for heat to flow out of the parcel as it moves over long distances, and energy is not lost to the surroundings. Thus if the parcel is compressed, its energy density increases: the fluid gets hotter. This increase in temperature with depth in the core is estimated to be rather small, about .8 degree C. per kilometer. Nevertheless, the adiabatic gradient that would exist in the thermally convecting core would have a profound influence on the evolution of the mantle and crust above.



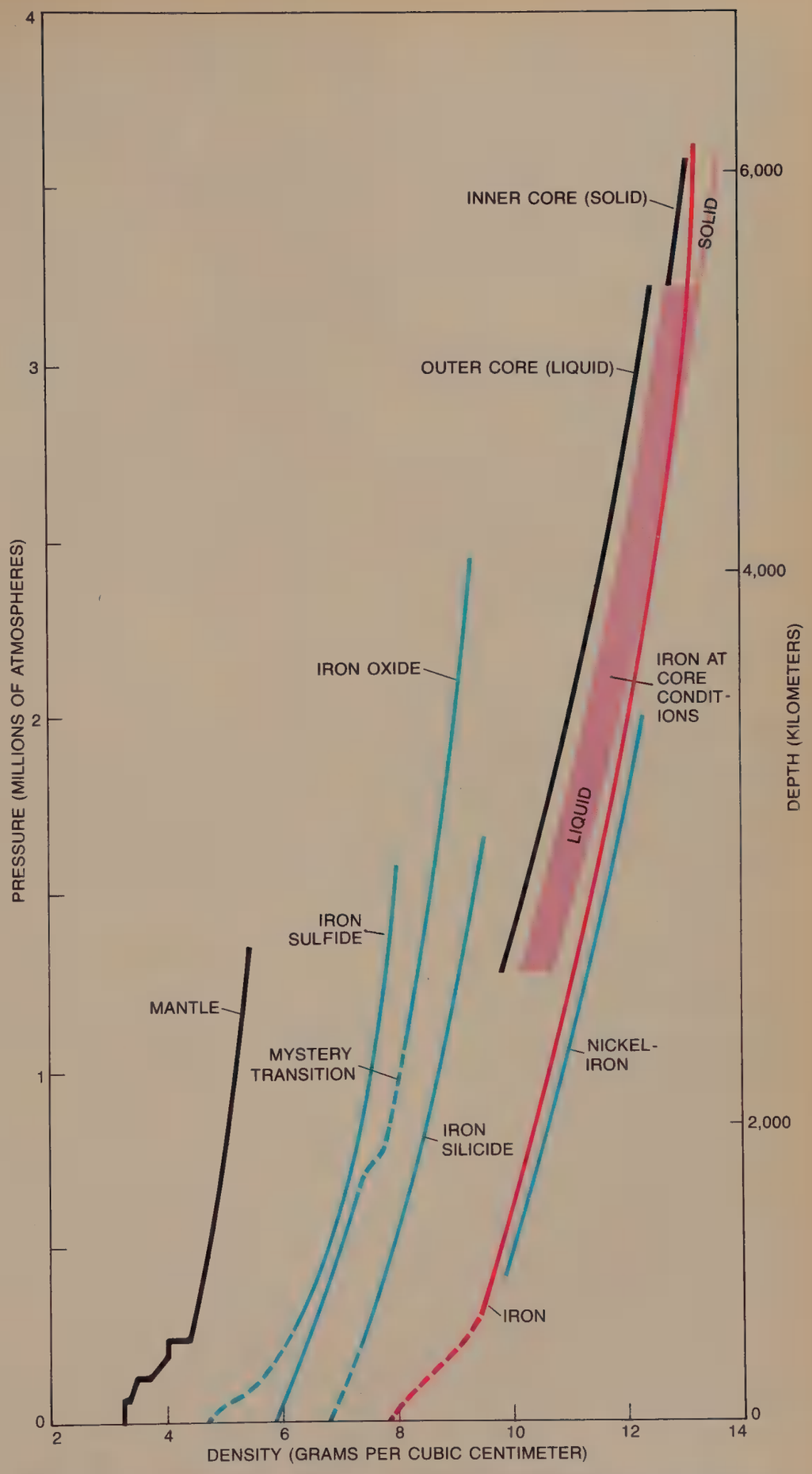
ACOUSTIC PROPERTIES OF THE CORE as revealed by changes in the velocity of seismic waves are shown as a function of depth from the surface of the earth and of pressure at each depth. One atmosphere is the pressure at the surface of the earth. The experimentally determined acoustic velocity of molten iron at core conditions (the pressures and temperatures of the core) closely matches the observed velocity. The seismic data also suggest the presence of an anomalous zone (*vertical shading*) at the top of the inner core. The zone is characterized by a relatively high attenuation of seismic waves. This is thought to be a partially molten region.

According to Fourier's law, heat is conducted down a temperature gradient with a flux (thermal energy passing through a surface of unit area in unit time) given by the value of the temperature gradient times the thermal conductivity. No one has measured the thermal conductivity of iron alloys at the temperatures and pressures of the core, but R. N. Keeler and G. Matassov have determined the electrical conductivity of alloys at core conditions by means of shock-wave experiments done at the Lawrence Livermore National Laboratory. In these metals the electrons carry the thermal energy as well as being responsible for the electrical conductivity. Thus one can employ the data to estimate the thermal conductivity of the outermost core; values between about 60 and 110 watts per degree C. per meter seem appropriate. Multiplying them by the adiabatic gradient of .8 degree per kilometer, one finds the surprising result that the predicted heat flux out of the core (70 milliwatts per square meter, plus or minus 25 milliwatts) is the same as the average heat flux at the surface of the earth.

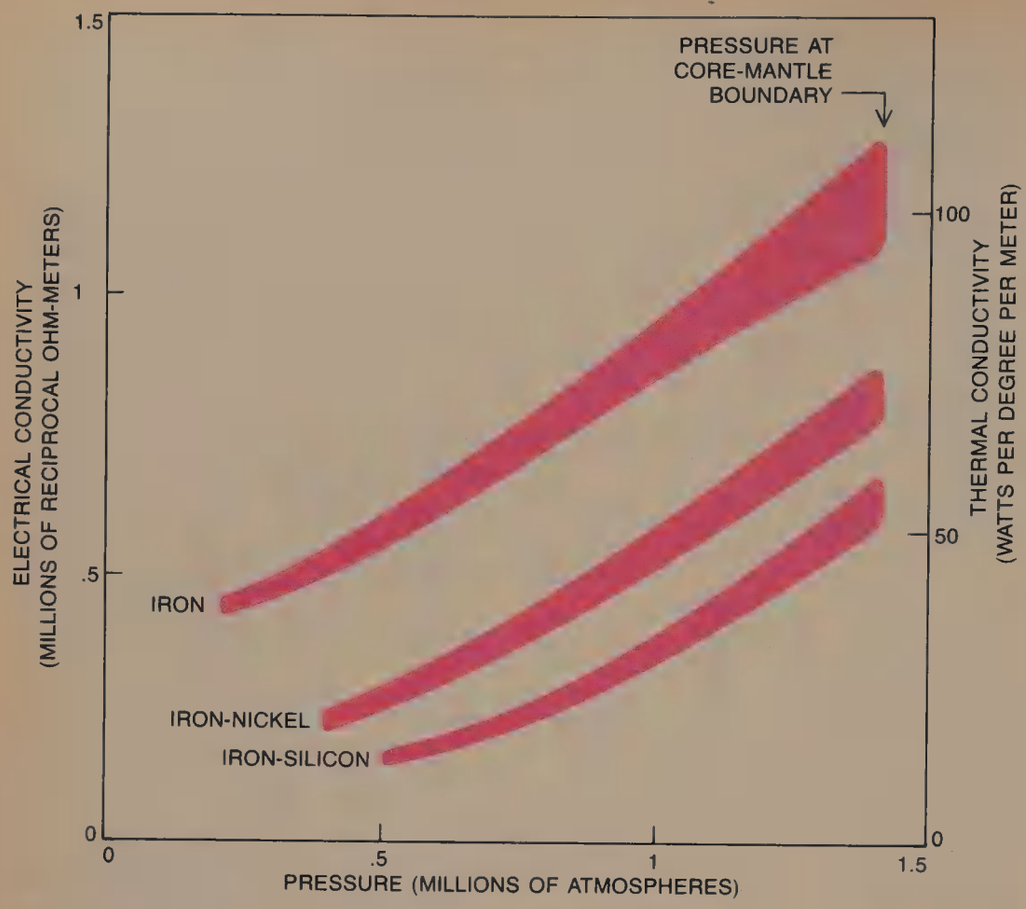
In fact, as most investigators currently believe, if thermal convection occurs, one would expect much more heat to be transported with the fluid than conducted. Therefore the heat flux from the core into the mantle would be higher still. The core thus becomes one of the major sources of heat driving convection in the mantle. Hence there may be not only an indirect but important link between the dynamics of the core and the earth's magnetic field but also one between the dynamics of the core and the large-scale tectonic motions observed at the earth's surface. Alternatively, lower temperature gradients in the core and a lower heat flux into the mantle are possible to the extent that the dynamo is compositionally driven. A low heat flux would imply the core exerts little influence on the dynamics of the mantle.

If a thermally driven geodynamo would need no compositional difference, a compositionally driven one would need no thermal difference. Therefore if no heat sources are invoked and the convective motions in the outer core are assumed to originate in the separation of dense crystals from the core fluid, the core could theoretically be isothermal: it would show essentially no change in temperature with depth. This is an extreme case, and it seems more plausible that heat sources are at work in addition to the compositional mechanism. Hence even for the compositionally driven dynamo temperature would be expected to increase with depth in the core. The core must still be considered an important source of heat for the overlying mantle.

The compositional dynamo does have



DENSITY AS A FUNCTION OF PRESSURE is charted in a comparison of the observed values for the earth's mantle and core with experimental data on iron and iron alloys that may exist in the core. The shock-wave experiments were done at the California Institute of Technology and at the Los Alamos National Laboratory. The comparison is improved by correcting the experimental data for the temperature within the earth and for the liquidity of the outer core, as is shown for iron by the shaded band. The width of the band corresponds to the uncertainty in density at each pressure. The "mystery transition" in iron oxide is a rapid increase in density with pressure that is seen in the experiments. The reason for this densification is not known, but it may indicate that the oxide becomes metallic at core pressures. If this is the case, oxygen may be the alloying element that modifies the observed density of iron in the core.



ELECTRICAL CONDUCTIVITY of iron and two iron alloys at high pressure is shown on the basis of shock-loading experiments. The data are corrected to a uniform temperature of 3,200 degrees Celsius to approximate conditions in the core. Because heat is transported by the electrons in metals, the data make it possible to estimate the thermal conductivity in the core. Near the top of the core it is evidently from 50 to 70 watts per degree per meter for the alloys, 10 times higher than in the overlying mantle. The finding suggests that the flux of heat from the core to the mantle makes the core a major source of heat driving convection in the mantle.

the simple requirement that the liquid of the outer core can separate into two phases (presumably solid and liquid) of significantly different composition, so that they can have significantly different densities. After separation the remaining liquid must be less dense than the original liquid, which makes up the rest of the outer core.

A liquid with a composition different from that of a solid is exactly what would be expected for a partially frozen alloy under equilibrium conditions. This is why an alloy melts and freezes at slightly different temperatures. Here, of course, the presumption is that the solid inner and liquid outer parts of the core are at equilibrium and that accordingly they differ in composition. Certainly the presence of a seismic attenuating zone at the top of the inner core—possible seismological evidence for a crystal-liquid mush—supports the idea.

On the other hand, one should bear in mind that the magnetization of rocks suggests the basic structure of the core (for example the presence of an inner core) may have been unchanged for most of geologic history. If the geodynamo is powered by the separation of the inner core from the outer one, it may be difficult to avoid accepting the contradictory assertion that the structure of the core has changed drastically over geologic time.

A stronger test would be to see whether the inner core is significantly denser than what would be expected for solidified outer-core material. The present seismological evidence is not sufficient to answer this question unambiguously. What seismic waves show is that the jump in density across the boundary between the inner core and the outer one is quite similar to the change in density of iron solidifying at the same pressure without a change in composition. This finding does not support the compositional-dynamo model because it implies that the solidified part of the outer core has essentially the same density as the inner core. Geophysicists are tantalizingly close to being able to evaluate this model, but more data are needed from both seismology and experiments at high pressure.

I return now to the question of primordial heat because it bears on the questions of how and when the core formed and in what way the process was related to the birth of the planet. Two extreme scenarios can be identified. One is that the earth assembled first and then separated into distinct iron and silicate fractions: the core and the mantle. The other is that the core aggregated first and then the remaining silicate-rich material was added.

The first of these pictures is called homogeneous accretion. It is mechanically analogous to the compositional model of the geodynamo, entailing a separation of dense material from less dense material after the earth had accreted. This model is well entrenched in the geophysical literature.

The second picture is heterogeneous accretion. It is a more recent and somewhat less well defined model. The reason is that different rationales have been proposed for the core metal to accrete before the silicates of the mantle do.

In order to distinguish between models of accretion one must consider the timing of three separate (but not necessarily temporally distinct) events in the earliest history of the solar system: (1) the condensation of solids out of the gaseous, cooling solar nebula; (2) the accretion of the entire earth, and (3) the accretion or formation of the core. This is the order of events for homogeneous accretion, with the key factor in the model being that all solids are condensed before accretion begins. Thus the growing planet accumulates both silicate and metal at the same time.

Subsequently the core separates from the mantle. According to Francis Birch of Harvard University, the separation leads to the release of an enormous amount of gravitational energy as the dense iron settles to the center of the planet. The amount of energy involved is comparable to the total thermal energy that would leave the earth over 4.6 billion years, given the present heat flux at the surface. It would have been enough to heat the entire planet by a few thousand degrees, which would presumably initiate substantial melting.

The best-known model of heterogeneous-accretion, proposed by Karl K. Turekian and Sydney P. Clark, Jr., of Yale University, visualizes the core material as condensing early and accreting before condensation is complete, indeed before the mantle silicates can condense and begin to accrete. Hence by the time the mantle has accreted the core is already in place because of the prior chemical separation of iron and silicates during the condensation process. The result is that the planet is relatively cold once it has formed. The reason is that little heat is retained as the earth accretes small particles condensing out of the solar nebula; most of the heat can be efficiently radiated back into space.

With no heat released by the separation of the core and the mantle, high temperatures would not be reached in this model. Yet the chemical separation of iron and silicates during condensa-

tion is no longer thought to be plausible. Calculations of the temperatures at which minerals would condense out of the cooling solar nebula indicate that the mantle phases appeared at about the same time as the core materials and maybe earlier. Therefore chemical heterogeneous accretion is now considered to be an unacceptable model.

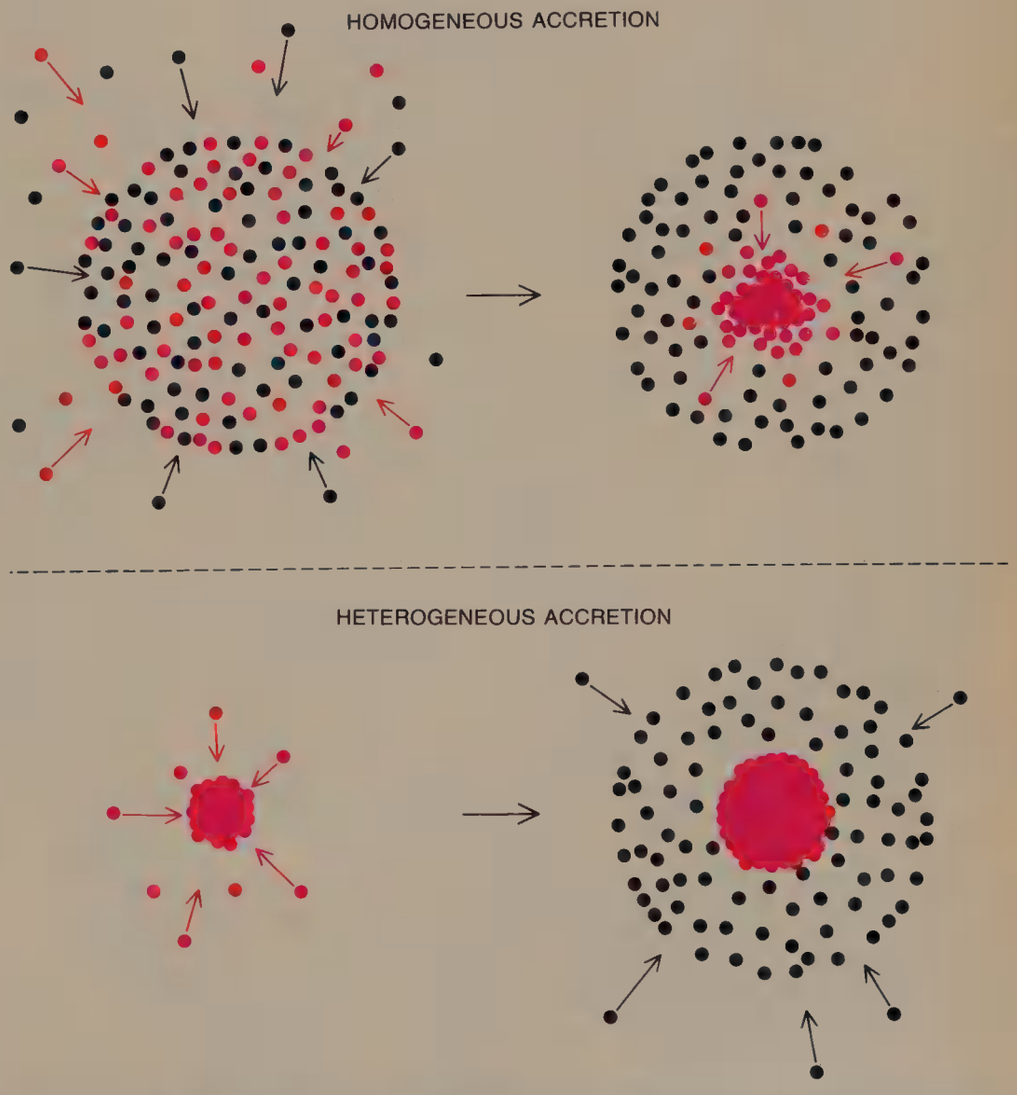
An alternative, physical mechanism could lead to heterogeneous accretion after condensation is complete. As proposed by Egon Orowan of M.I.T. and Hitoshi Mizutani and Takafumi Matsui of the University of Tokyo, after the materials of the core and the mantle have condensed one would expect metallic (core-forming) grains to accrete preferentially. In this view, because of the brittleness and rigidity of silicate and the relative ductility and high density of iron-rich phases, one would expect a rapid agglomeration of metal and a much slower accumulation of silicates. With increasing size the growing protoplanet can accumulate silicates more readily as the gravitational attraction increases. Thus most of the mantle could be accreted after the core has assembled. In this model one can think of planetary growth as being nucleated by the formation of a metallic core.

All the evidence suggests that the earth accreted after metal and silicate particles had condensed in the solar nebula. The model of homogeneous accretion and the model of physical heterogeneous accretion could accommodate this process. The present thinking is that not only particles but also small planetesimals, perhaps already differentiated into iron (core) and silicate (mantle) regions, may have been the building blocks of the earth. One line of evidence is that iron meteorites are thought to have been present 4.6 billion years ago. They exhibit textures characteristic of a slow cooling that could have occurred only in planetary bodies with dimensions of hundreds of kilometers or more. Protoplanetary bodies of substantial size therefore could well have been present at the time the earth formed. Computer studies show that if such bodies were accreted rapidly, the radiation of heat into space would have been relatively inefficient and the planet would have got hot as it grew.

Without rejecting the possibility that the nucleation of the earth may have been initiated by the accumulation of iron, one should consider what would have happened as the planet was growing. On the assumption that both silicate and iron were being accreted, as in the homogeneous-accretion hypothesis, one finds it is not possible to delay for long the settling of iron toward the center of the planet. As G. F. Davies of Washington University has recently pointed out, the force of gravity pulls increasingly on the denser iron-rich regions with time. The reason is that the gravitational force increases as the planet grows, which is to say as its mass increases. For reasonable estimates of the sizes of the denser regions Davies finds the iron can sink, without necessarily melting, after only about one-eighth of the final mass of the earth has been accumulated. The phenomenon is explained by the fact that rocks remain relatively weak even at the pressures existing deep inside a planet.

It therefore seems inevitable that the formation of the core began well before the earth was fully formed and that the differentiation of the planet took place at the same time as most of its accretion. This picture has features of both the homogeneous- and the heterogeneous-accretion models: the earth accretes after condensation is complete, but the core is present early in the growth of the planet. One implication is that the core probably started forming at a relatively low pressure. This implication may lead to difficulties with the hypothesis that the core is an alloy of iron and oxygen. Still, the relation between the composition of the core and the processes by which it formed is not well understood, and further work in this direction is needed.

In any case, once the differentiation of iron and silicates begins the planet would be expected to heat up rapidly as gravitational energy is released. This heating and the heating caused by the relatively rapid accretion of planetesimals are thought to be enough to trigger melting, which leads to an even more effective differentiation of the planet. In this way the formation of the core is seen as a self-perpetuating and possibly accelerating process. Evidently the core had a role in triggering the geologic processes that are still taking place, some 4.6 billion years later.



FORMATION OF THE EARTH is visualized as having been by one or the other of the processes depicted here. In the homogeneous-accretion model silicate (*black*) and iron (*color*) accumulate to form the complete planet (*top left*). Subsequently the core forms by the separation of the metal from the silicate (*top right*). During the formation of the core the iron sinks to the center of the planet and heat is generated by the release of gravitational energy. In the heterogeneous-accretion model the metallic core is accumulated first and the silicate mantle accretes around it. The sequence could occur during or after the condensation of solids out of the solar nebula, depending on whether chemical or physical processes are involved. In each model the accretion of the planet is viewed as resulting from the infall of meteoritic bodies.



The Earth's Mantle

The great shell of silicate that lies above the metallic core is heated by the decay of radioactive isotopes. The heat energizes massive convection currents in the upper 700 kilometers of the ductile rock

by D. P. McKenzie

Whereas the dynamics of the earth's core are reflected in the frozen polarity of ancient magnetic fields and in the luminous spectacle of the aurora, change within the mantle of the earth is recorded in the surface configuration of the oceans and the continents. Large-scale motions in the mantle take the form of currents of dense, solid, ductile material. The surface expression of these currents is the movement of the plates that form the solid surface of the earth, most of which include parts of both oceans and continents. The familiar map of the world is therefore a snapshot of dynamic processes that have their origins in the mantle. The snapshot does not, however, reveal much about the deeper circulation.

The major difference between continents and oceans is the thickness of their crust. The continental crust is anywhere from 20 to 70 kilometers thick; the mean oceanic crust is only about six kilometers thick. The plates therefore have to be at least 70 kilometers thick if they are to carry continents. It is now clear that the thickness of the plates actually has nothing to do with the thickness of the crust, and that all plates are about 100 kilometers thick, regardless of whether they are covered by continental or oceanic crust. This layer from which the plates are made is called the lithosphere.

The currents that move the plates are convection currents, in which hot material rises and cold material sinks. New plate is formed at a midocean ridge and cools as it moves away from the ridge toward an oceanic trench. The plate is then subducted by the trench and sinks into the mantle, because it is colder and denser than the surrounding material. As the dense plate sinks it releases gravitational energy and carries cold material downward, forming a sinking cold limb of a convection cell. These two parts of the convective circulation, namely subduction and the motions of the plates at the surface, can be directly observed.

In addition there must be a horizontal return flow that carries material from the trench back to the ridge, and an upward movement of hot material to balance the downward movement near trenches. Nothing can be learned about this part of the circulation from the movement of the plates. In particular there is no reason to believe the upward flow coincides with the position of a midocean ridge; indeed, it is now known it does not. The motion of the plates also tells us nothing about any mantle flow whose horizontal extent is considerably smaller than the size of the plates themselves. The reason is that the plates move rigidly, and their motion responds only to the average of the forces on their base and their boundaries. Clearly little can be learned about mantle convection from past and present plate motions, even though these motions are now known in great detail. This situation was thoroughly unsatisfactory: everyone believed convection in the mantle provided the energy to maintain the motions of the plates but no one could explain in any detail how it did so!

Fortunately the behavior of convecting fluids had been studied in considerable detail by workers interested in fluid dynamics, who had developed a variety of methods for modeling such flows. The most straightforward models consist of a layer of fluid in a shallow tank heated from below or from within. The behavior of the fluid can be studied by measuring the temperature within it or by shining light vertically through it; the light is focused by cold, sinking sheets to form bright lines. Another method of investigating convection, which Nigel O. Weiss and I have been pursuing intensively at the University of Cambridge, is to solve the relevant equations numerically on a large computer. The two methods are in many ways complementary. The tank experiments can easily be done in three dimensions, but they are little help in understanding geophysical observations. The numerical experiments are at present limited to two-dimensional flows, but they can serve to relate the observations to the convective circulation. In carrying out such modeling our aim was to understand the physical processes that control convection, not to create a mathematical model of the mantle.

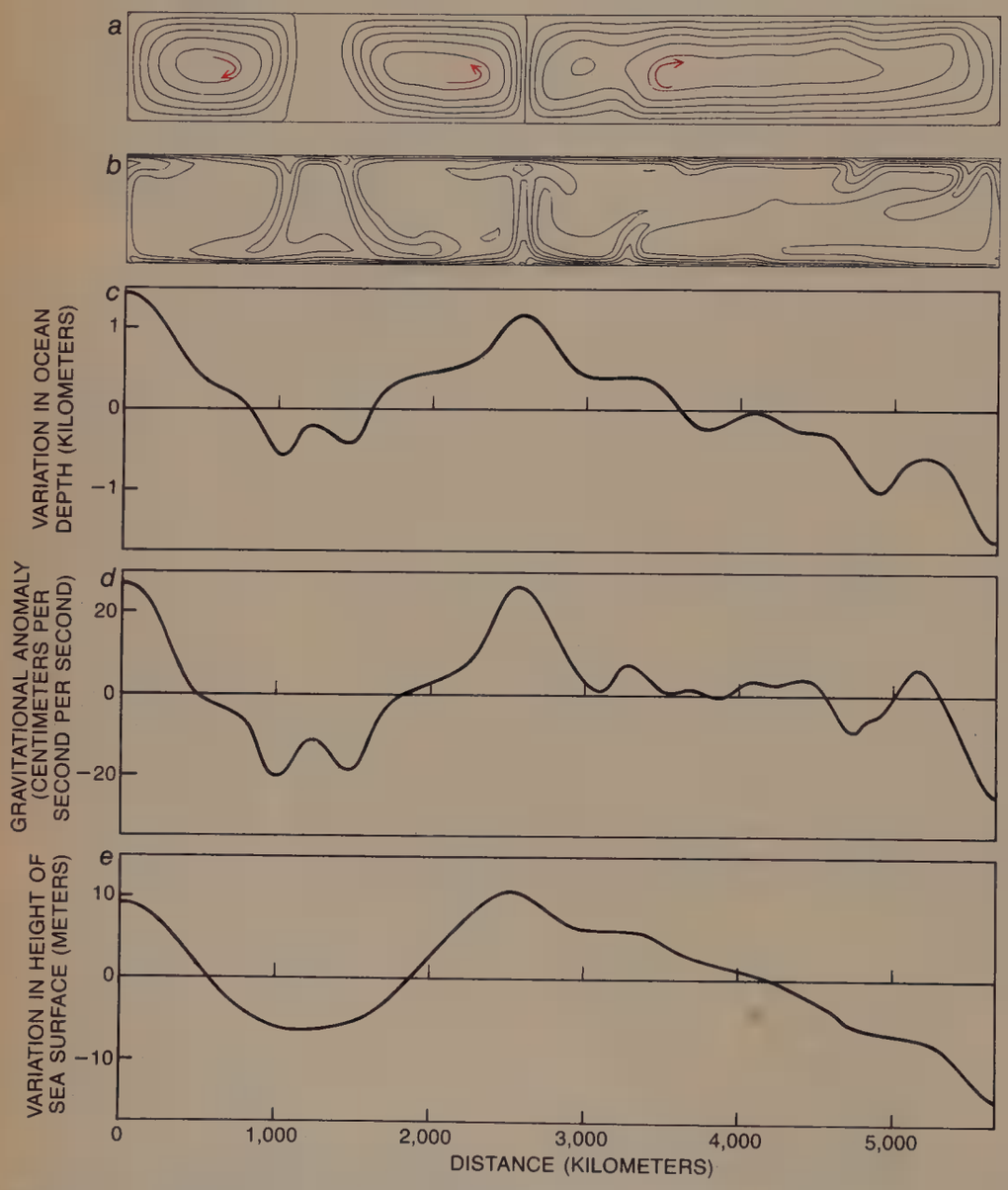
Both the tank and the numerical experiments produced convection cells in which the separation between cold, sinking regions was about twice the depth of the convecting layer. Tank experiments with a moving upper boundary showed that the convection could then occur in rolls parallel to the direction of motion of the boundary. The convecting layer in the earth was thought to be about 700 kilometers thick, and so these fluid-

WINDOW ON THE EARTH'S MANTLE is afforded by the island of Hawaii, which has been built up entirely by volcanic eruptions. Two major volcanoes make up the island; Mauna Kea, in the upper part of the photograph on the opposite page, is now dormant, but Mauna Loa, in the lower part, has erupted several times in this century. Older, weathered lava radiates in brown bands from the summit of Mauna Loa. Darker radial bands mark relatively recent lava flows. The major site of current volcanic activity, however, is along the southeastern flank of Mauna Loa at the Kilauea caldera and near the coastline southeast of Kilauea. The chronological trend is therefore from the northwest to southeast; indeed, along the entire Hawaiian ridge the seamounts and islands become younger to the southeast. The oldest parts of the ridge are the Emperor seamounts in the northwest Pacific, and the youngest island is the one in the photograph. The lava emerges at the center of a hot, upwelling region in the mantle, which is much larger than the Hawaiian ridge itself. The ridge is thought to record the slow northwesterly movement of the Pacific plate with respect to the upwelling region. Isotope measurements on Hawaiian lava show that the material has not been well mixed with the rest of the upper mantle and must be derived instead from a part of the mantle that has been isolated from the convecting parts of the upper mantle for perhaps a billion years. The lower mantle is a possible source. The photograph is a digital reconstruction of two images made by Landsat; the reconstruction was done by the Flagstaff Image Processing Facility of the U.S. Geological Survey.

dynamical experiments suggested that mantle convection consisted of at least two scales of motion. A small-scale circulation with a distance of about 1,500 kilometers between cold, sinking regions should be superimposed on the larger-scale circulation that returns material from a trench to a ridge. This two-scale model of circulation in the mantle could reconcile the geophysical observations with the behavior of convecting fluids, but the existence of the small-scale circulation could not be directly observed until recently.

The key to mapping the small-scale flow came with the introduction of remote-sensing devices that have made it possible to map the earth's gravity field with great accuracy. The numerical calculations had shown that upwelling regions of convection should be associated with small, positive gravitational anomalies and that flow should push up the earth's surface. Plates are too thin to have much effect on either the gravity field or the surface deformation; therefore if the gravity field and the surface deformation can be accurately mapped, it should be possible to see through the plates and map the convective circulation under them.

In practice it is not the gravity field but the shape of the sea surface that can be measured from a satellite. Since the water is attracted to regions where the gravity field is stronger, there is a close relation between variations in gravity and the height of the sea surface. In this way Barry Parsons of the Massachusetts Institute of Technology, Anthony Watts of the Lamont-Doherty Geological Observatory of Columbia University, Micheline Roufosse of the Center for Astrophysics of the Harvard College Observatory and the Smithsonian Astrophysical Observatory and I have been able to map the small-scale convection cells in the mantle. The maps confirm the general features of the two-scale model, but they leave two major questions unresolved: How does the flow evolve, and to what depth does the circulation extend?



NUMERICAL SIMULATION OF CONVECTION in the mantle shows that the flow is complicated and varies with time. The simulation is carried out by assuming the mantle is a two-dimensional fluid layer of constant viscosity, uniformly heated from below, cooled from above and thermally insulated on both sides. Heat is carried through the fluid by convection; the heated material moves upward and the cooled material moves down. The circulation of the material in the cells is shown at the top (*a*); temperature contours at intervals of 100 degrees Celsius record a rising plume of heated material near the center of the fluid layer and regions of colder, sinking material at the left and right (*b*). The model can also predict certain observable effects of the convection, such as the variation in the depth of the ocean (*c*) or the variation in the local acceleration of gravity, which is termed the gravitational anomaly (*d*). An important new method for measuring the gravitational anomaly is to determine the variation in the height of the sea surface caused by the local force of gravity. The model can also predict such variation (*e*).

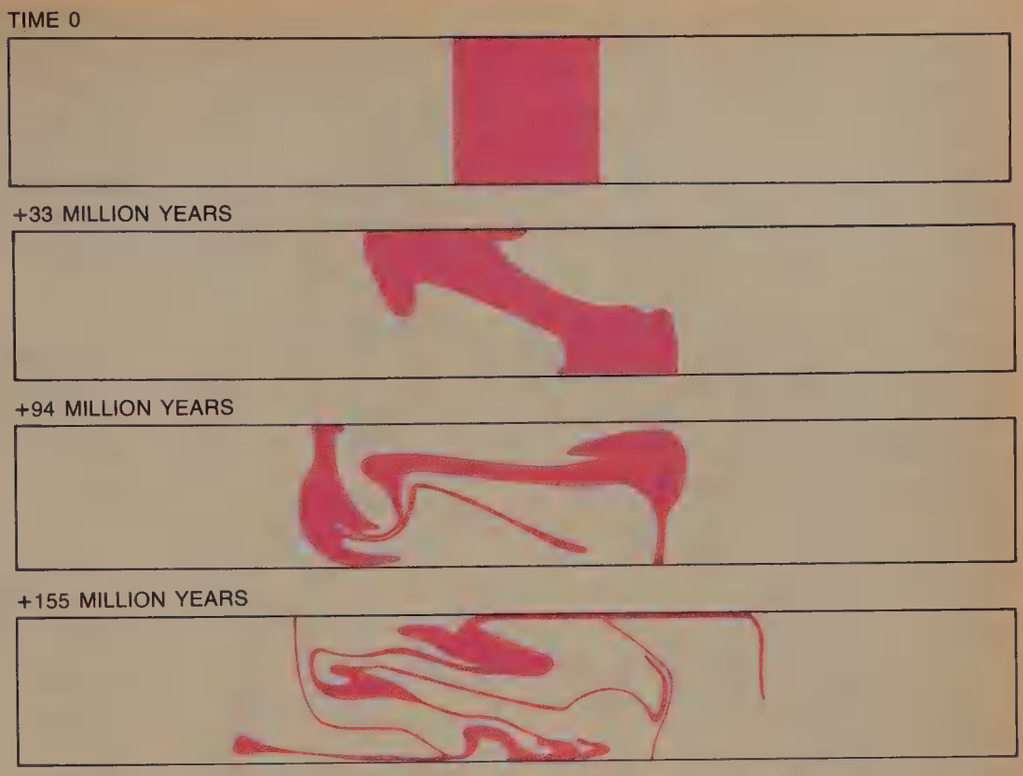
Before attempting to answer these two questions in detail it is worth noting they take for granted that the mantle can flow like a liquid. This behavior is quite unlike that of familiar crystalline solids at room temperature. Such solids generally deform elastically under small stresses, and they only flow when they are subjected to stresses in excess of their yield point. The yield point of mantle rocks is very large, much greater than any estimate of the stresses in the mantle, and so many geophysicists did not believe the mantle could flow. The solution to the problem became apparent when the flow behavior of materials close to their melting point was examined. Such studies were required in order to predict the behavior of materials at high temperatures, for instance those encountered in jet engines and in nuclear reactors. It was found that under these conditions all crystalline materials flow under any stress, however small it may be. The process is called creep, and it is familiar at low temperatures. High-temperature creep, however, differs in many important ways from low-temperature creep.

The creep of solid rock at high temperatures has resolved an important puzzle in geology. Observations made in the 19th century clearly showed that mountains, in spite of their great mass, are not associated with a positive gravitational anomaly. The excess mass above sea level must therefore be compensated for by low-density roots thrust into the mantle. Geologists had recently discovered that major mountain belts were raised by the thrusting of sheets of continental crust over each other, and so it was natural to argue that the same process could produce the low-density roots. The entire mass floats in the denser mantle, as icebergs do in water. The process is known as isostatic compensation. Furthermore, it was widely believed the lithosphere was everywhere underlain by a dense molten layer from which volcanic magma came, and it seemed entirely reasonable that the flow associated with isostatic compensation occurred within this layer.

Such evidence that the mantle could flow was largely neglected, however,

when seismologists began to probe the structure of the mantle by analyzing the waves generated by earthquakes. They found that *S* waves, or waves oscillating at right angles to the direction in which they propagate, can travel almost everywhere in the mantle. Such waves cannot pass through the body of a liquid, and so the 19th-century idea that the lithosphere floats on a molten layer of magma had to be abandoned. It was not until the discovery of sea-floor spreading in the late 1960's and the recognition that this newly created oceanic lithosphere eventually cools and sinks once again into the mantle that the seismological evidence was reconciled with the presence of mantle flow. Indeed, if high-temperature creep had been widely recognized in 1910, when Alfred Wegener first proposed that the continents drift, the response to Wegener's ideas might have been more favorable than it was.

Ironically, some of the most convincing evidence for flow within the mantle has now been derived from seismic studies. The foci of earthquakes under island arcs define a sinking slab whose thickness is rarely greater than 50 kilometers, and that dips from the oceanic trench at the surface at an angle of between 30 and 90 degrees. The shape of the slab is controlled by the shape of the island arc, which overrides the slab and shears it like the cutting tool of a lathe. Thus the data from seismic studies show a close connection between material near the trenches and material as deep as the earthquake foci: about 700 kilometers.



CONVECTIVE STIRRING IN THE MANTLE is vigorous enough to mix thoroughly the material in the upper mantle within several hundred million years. The stirring can therefore account for a surprising uniformity of the basalts collected along midocean ridges throughout the world: all such basalts include certain trace elements, such as neodymium, whose isotopes are found in a constant ratio to one another. Here a numerical model similar to the one on the opposite page is used to follow the evolution of a square patch of fluid whose properties are identical with those of the rest of the fluid in the convecting layer. The material in the patch is deformed into thin sheets on a time scale that is short compared with the time necessary to generate measurable variations in the isotopic ratios among the trace elements. This study of mixing in the mantle was carried out by Nicholas Hoffman of the University of Cambridge.

According to the theory of plate tectonics, the slabs are large, cold layers of mantle material that form the lower part of a plate. The slabs sink because they are denser than the hotter material that surrounds them. As they move downward and away from the trench, they tend to fracture vertically, and the part of the slab below the vertical crack tends to slip even farther downward. The sudden release of energy from the slippage generates an earthquake; this kind of earthquake is not observed, however, below 300 kilometers. On the other hand, at depths greater than 600 kilometers the part of the slab below a vertical fracture slips upward, as if the tip of the slab had met with considerable resistance to its motion. Meanwhile the entire slab is heated, and it gradually becomes hot enough to creep and so ceases to generate earthquakes. The process takes about 10 million years.

The mapping of the path of a sinking slab was one of the first empirical confirmations that convection currents in the mantle exist. As I have emphasized, the descent of a large quantity of cold material must be balanced by upward movement elsewhere in the mantle, and the transport of material from midocean ridges to trenches must be matched by deeper movements in the opposite direction. In order to understand such convective motion Nigel Weiss and I began 15 years ago to construct two-dimensional numerical models, based on the best contemporary estimates of the temperature, density and viscosity of the mantle, that we hoped would clarify the main features of convective flow.

In the model we assumed that the mantle is a two-dimensional object, rectangular in shape and 700 kilometers deep. At the bottom of the rectangle a constant flux of heat energy is introduced, representing the input of heat from the core of the earth and from the deeper part of the mantle. The sides of the rectangle are considered perfect insulators, and so heat can emerge from the rectangle only from the top, corresponding to heat lost through the earth's surface. Inside the rectangle there is a fluid whose viscosity is constant and whose density varies only with temperature. The calculation is done on a computer by dividing the rectangle into many smaller rectangular cells and then calculating the temperature, the density and other properties of the fluid in each cell at a given time. The effects of each cell on its neighbors are then extrapolated over a short time increment, and the values for the properties of the fluid in each cell are recalculated. By repeating the procedure over many increments of time the evolution of the fluid circulation throughout the rectangle can be approximated to a high accuracy.

Although the model is obviously much simpler than the real mantle, the flow within the rectangle is complex and varies with time [*see illustration on opposite page*]. Several small-scale cells are generated with variable size and shape. No cells on a scale large enough to account for the plate motions are generated in this experiment (although they were in others). Upwelling regions of hot material do not remain fixed; they shift with respect to one another at a rate equivalent to about a centimeter per year in the real mantle. The overlying lithosphere would be pushed upward above an upwelling region and pulled downward above a downwelling region by as much as a kilometer. The gravitational anomaly and the associated variation in the shape of the sea surface are calculated for each point along the upper edge of the rectangle.

The calculation of the gravitational anomaly is complicated. At first it might seem that the anomaly above a hot, rising region of a convection cell would be negative, because the density of the hot material is less than that of the cooler surrounding material. The elevation of the surface above an upwelling region,

however, generates a positive anomaly because of its additional mass. The two contributions tend to cancel, but the positive contribution from the surface deformation is slightly larger than the negative one from the hot material. Hence according to the model a positive gravitational anomaly should be found over the rising region of a convection cell and a negative anomaly should be found over a sinking region. In order to detect convection cells one should look for a pattern of positive and negative gravitational anomalies that is correlated with a similar pattern of anomalies in the elevation of the sea floor.

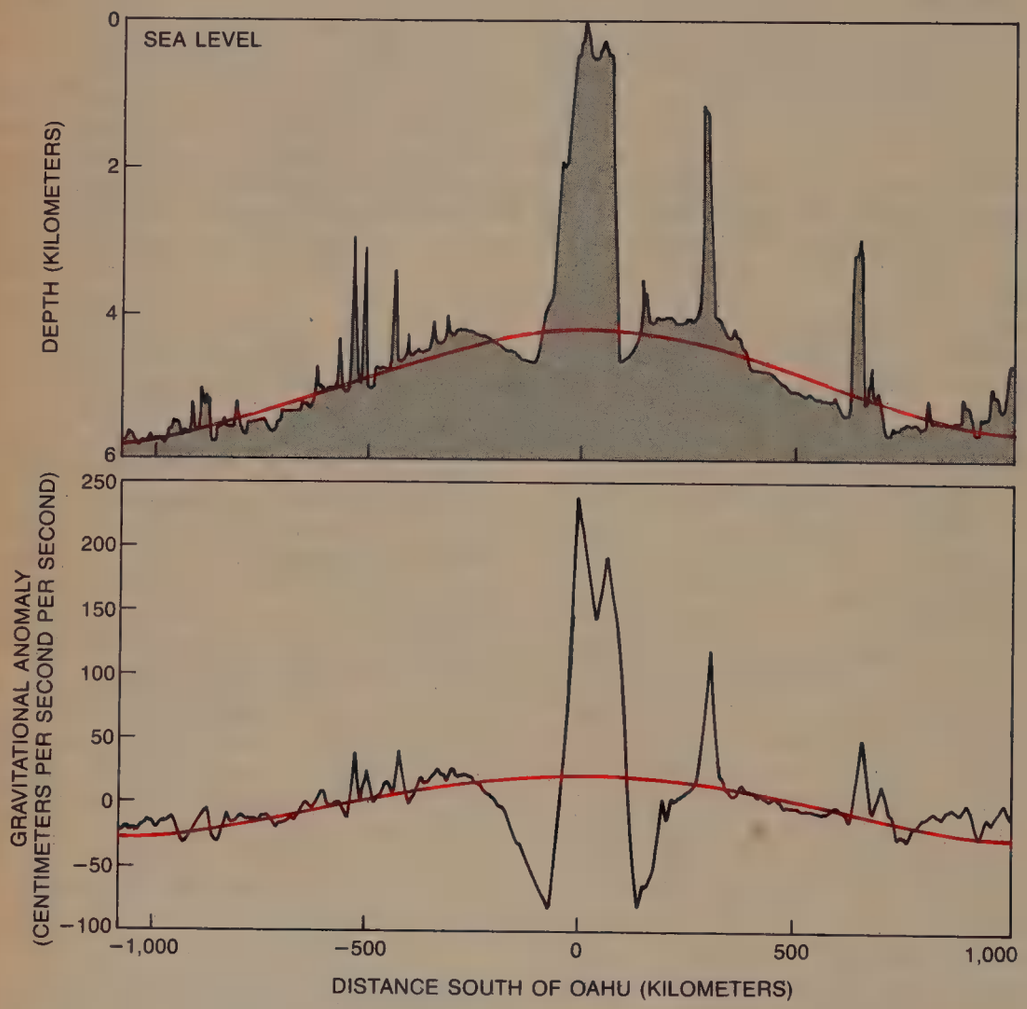
In 1978 Parsons, Watts, Roufosse and I began to analyze all available gravitational and bathymetric data in order to detect such a correlation and thereby map the circulation of the mantle. Both kinds of data present special problems that must be overcome before useful correlations can be made. The conventional method of measuring gravitational anomalies is to measure the extension of a spring in a delicate balance. The weight of the mass at the end of the spring varies with changes in the gravitational field, and so the extension of the spring gives the anomaly. The instrument works well on land but is less accurate at sea. Indeed, although we needed an accuracy of one part per million, we were not confident that the accuracy of measurements made aboard ships was better than one part per 100,000.

Remote-sensing methods offered an alternative. Seawater is attracted to regions where the gravitational anomaly is positive and forms a measurable bulge on the surface; a trough in the sea surface is formed where the anomaly is negative. The numerical model predicts that the change in elevation from bulge to trough caused by circulation in the mantle is about 20 meters over a distance of 2,000 kilometers. Although the variation is small, the effect can readily be measured with the aid of a satellite. At least two satellites, *Geos 3*, which was launched in 1978, and *Seasat*, which operated for three months in the same year before it failed prematurely, have made measurements of the height of the sea surface with radar altimeters. If the position of the satellite is known, the time radar waves take to travel from the satellite to the sea surface and back can be measured; the height of the sea surface can thereby be determined.

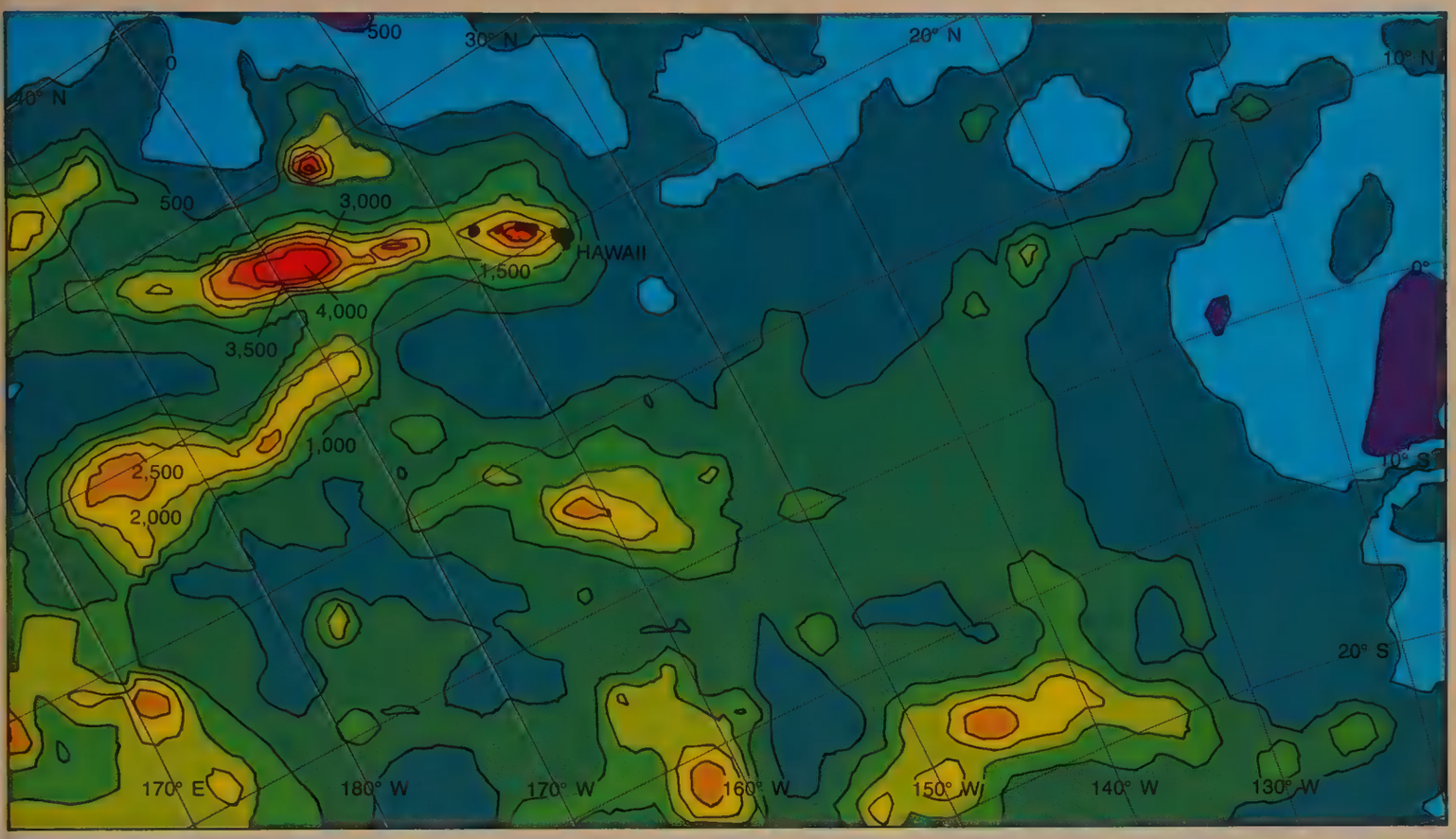
In order to correlate the altimetric measurements with the changes in the shape of the sea floor the depth of the oceans must be measured as well. All detailed bathymetry must still be done by measuring the interval for sound waves to travel from a ship to the sea floor and back again. Raw bathymetric data, however, cannot be used directly to map swells and depressions in the sea floor caused by the circulation of the mantle. Short-wavelength changes in depth (changes that take place over relatively short distances) are caused by elastic forces within the plates themselves and have nothing to do with the mantle. For example, when a load is placed on the surface of a plate, the plate bends downward. The most obvious causes of plate loading are volcanic eruptions, such as the ones still building the island of Hawaii.

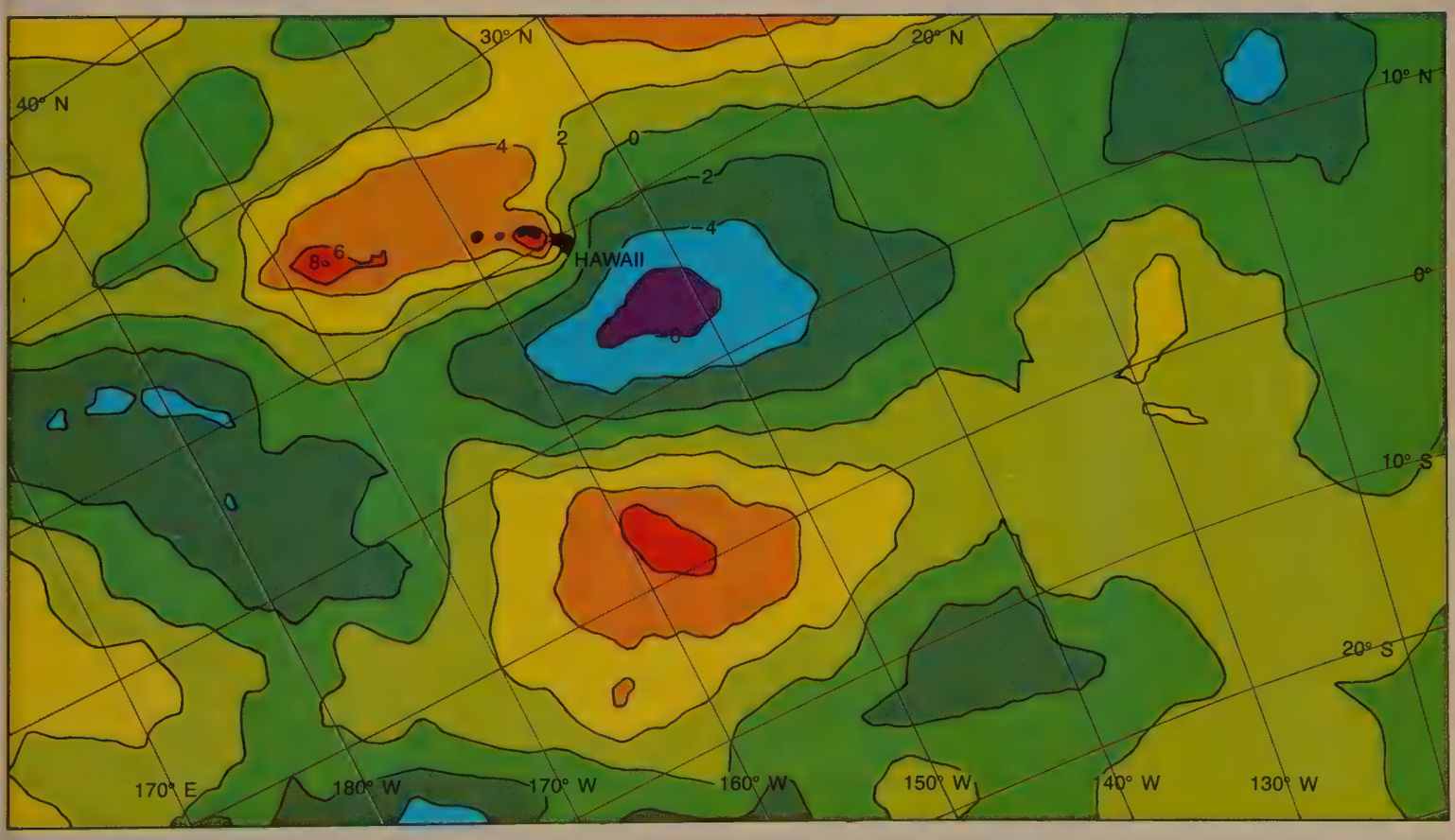
Watts has found that the bending of the plate caused by the weight of the volcanic rock making up the island and its submarine part is confined to within less than 200 kilometers of the load. By measuring the effects of plate loading in the vicinity of many islands and seamounts Watts has estimated that oceanic plates behave as if they were elastic plates whose thickness is 35 kilometers or less. Such a plate is too thin and flexible for its elastic response to local forces to affect any bathymetry or gravitational anomaly with a wavelength of 500 kilometers or more. Thus in our study of convection in the mantle we removed all variations in the depth of the sea and in the height of the sea surface whose wavelength is shorter than 300 kilometers.

Another important effect of the plates on the depth of the sea floor is the age of material in the plate. Midocean ridges are relatively shallow because the plate is hot and has a relatively low density. As the plate spreads away from the ridge and cools it contracts; the depth of the ocean therefore increases with the age of the plate at the bottom. By dating the sea floor on the basis of magnetic anomalies associated with reversals of the earth's magnetic field, the depth of the sea caused by the contraction of the plate can be estimated. Since the contraction is unrelated to the swells and depressions caused by circulation in the

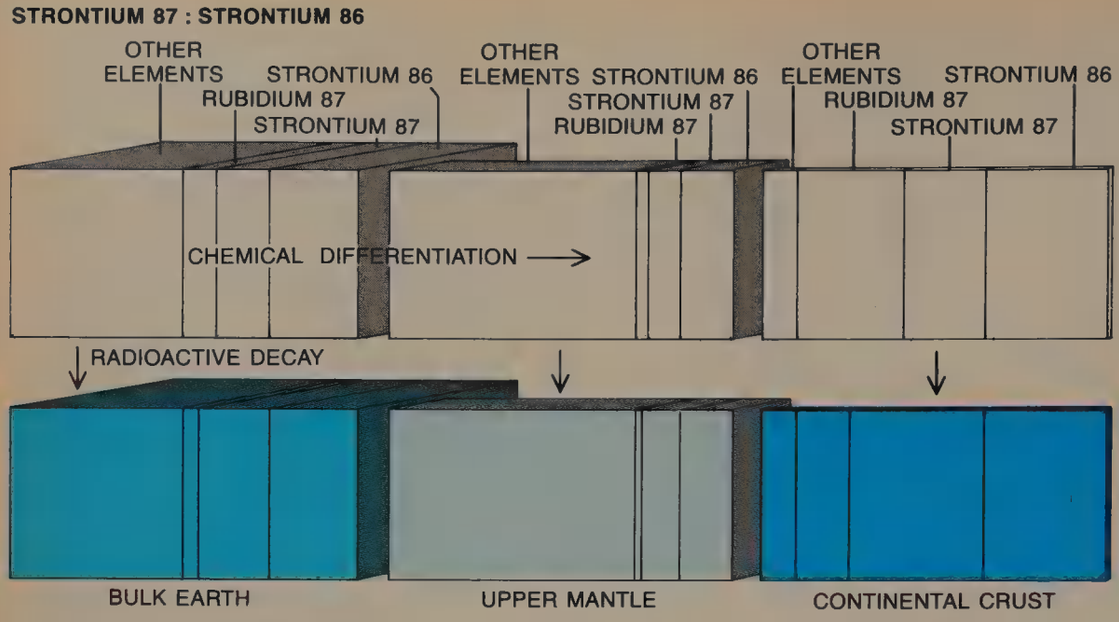


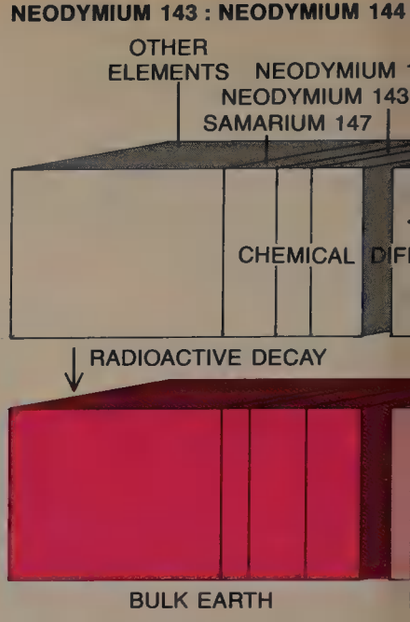
ELASTIC RESPONSE OF A PLATE to stresses that are unrelated to convection in the mantle must be understood before the effects of upwelling or downwelling in the mantle can be observed. For example, the weight of the Hawaiian ridge causes a depression of the sea floor on both flanks of the ridge, which can be seen in the profiles of both the bathymetry and the gravitational anomaly north and south of the island of Oahu. The deflection caused by the ridge is confined to within about 200 kilometers of the load. Hence it can be inferred that the rigidity of the plate is not great enough for local loads to generate variations in depth or gravity at greater distances. Both profiles also show that the Hawaiian ridge and the neighboring depressions are superposed on a broad swell, which is generally thought to be the surface expression of a hot, upwelling region in the mantle. The swell can be best observed if fluctuations in the bathymetry and gravitational anomaly shorter than 500 kilometers are smoothed (*colored curves*).



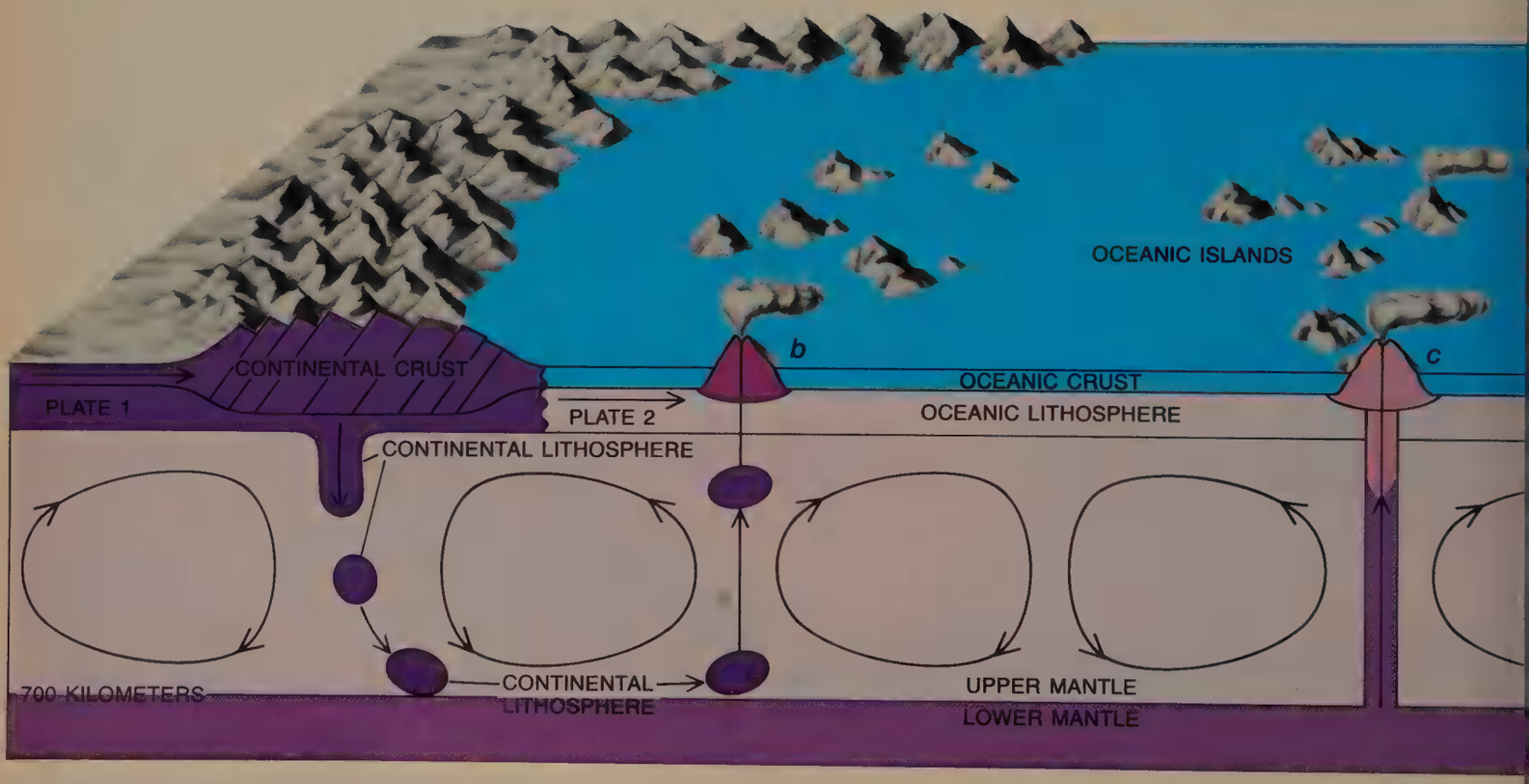


CONVECTION CURRENTS IN THE MANTLE under the Pacific plate are mapped by correlating bathymetric and gravitational anomalies. Both maps have been smoothed to eliminate fluctuations shorter than 500 kilometers; in addition the bathymetric anomaly is plotted as the so-called residual depth, which is the difference between the depth that can be attributed to the contraction of the oceanic plate as it cools and the observed depth (*upper map*). The gravitational anomaly is observed as a fluctuation in the height of the sea surface, which is measured by radar altimeters carried aboard satellites (*lower map*). The maps show that where the sea surface tends to bulge the residual depth is positive, and both features are expected above an upwelling region in the mantle. By the same token the sea surface tends to be depressed where the residual depth is negative, which is expected above a downwelling region. The map is projected in such a way that the motion of the plate with respect to the mantle is always to the left over the entire region. The motion generates a slight but detectable elongation of the anomalies in the direction of the motion, and so they resemble ellipses whose long axes run across the page.

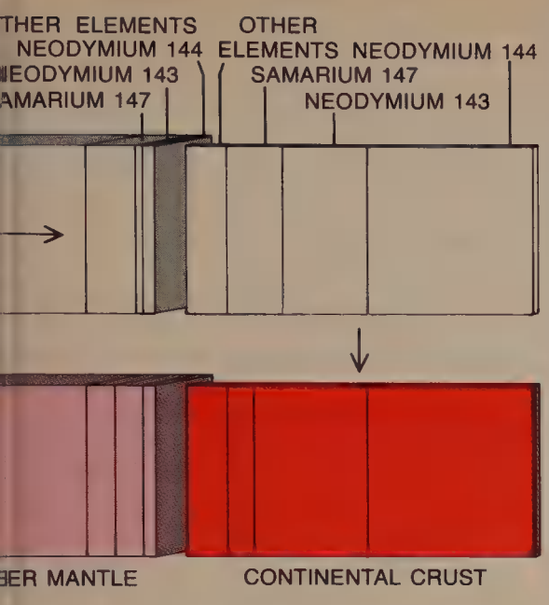




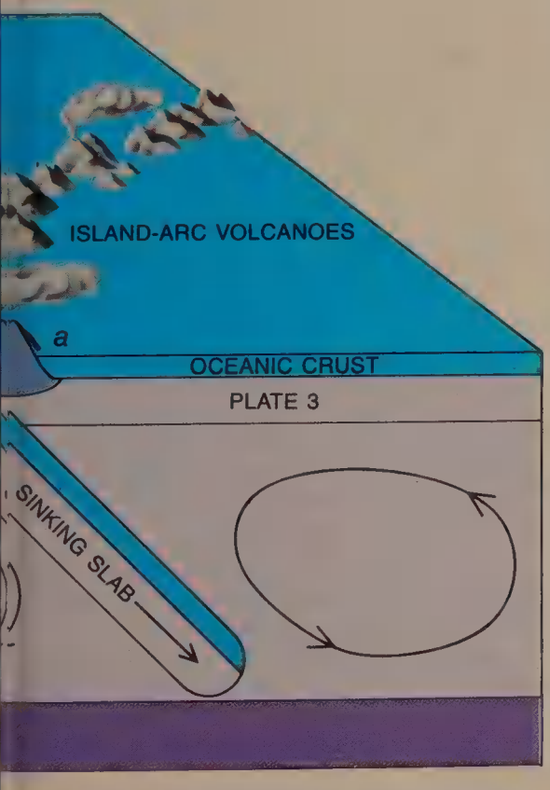
RATIOS AMONG THE ISOTOPES of an element can indelibly mark materials as having undergone chemical differentiation in the distant past, even though the ratios are not affected by recent chemical change. For example, in the early stages of the earth's history certain elements such as strontium and rubidium, whose ionic radii are relatively large, were concentrated in the crust of the earth and depleted in the upper mantle because their ions do not fit easily into the lattice framework of common minerals. The relative increase in the crustal concentration of rubidium was greater than that of strontium. In the schematic diagram the ratio of the isotope strontium 87 to strontium 86 remains the same after the differentiation of the crust and the upper mantle because all the isotopes of any element are chemically equivalent. The ratio of rubidium to strontium varies; thereafter half of the rubidium 87 in each part of the earth decays to strontium 87. (The actual fraction that decays is smaller, but here it has been exaggerated for the purpose of clarity.) Hence because of the initial surplus of rubidium 87 in the crust, the ratio of strontium 87 to strontium 86 grows most rapidly in the crust, less rapidly in the earth as a whole and least rapidly in the upper mantle. The shade of blue in the diagram indicates the present isotopic ratio: the darker the blue, the greater the ratio of strontium 87 to strontium 86. Similarly, the elements samarium and neodymium were both concentrated in the crust, but relatively more neodymium than samarium was transferred to the crust. Therefore the ratio of samarium to neodymi-



ORIGIN OF MATERIAL erupted onto volcanic islands can be inferred from the isotopic ratios of certain elements. The ratios are indicated by the color coding adopted in the illustration at the top of these two pages: material from the continental crust has a high ratio of strontium 87 to strontium 86 (*dark blue*) and a low ratio of neodymium 143 to neodymium 144 (*dark red*), and so it is colored dark purple. Similarly, material from the upper mantle is colored pale purple, and material from the lower mantle, whose isotopic ratios are assumed to be the same as those for the earth as a whole, is colored an intermediate shade of purple. Rivers carry large amounts of strontium into the oceans but only small amounts of neodymium. The ratio of strontium 87 to strontium 86 in seawater is therefore similar to that in the continents, and so it is colored dark blue. As seawater circulates through the oceanic crust it deposits continental strontium, and so the oceanic crust is also colored dark blue. Three possible origins of the material that forms volcanic islands are shown. Island-arc volcanoes are formed from a mixture of melt from the upper mantle and the oceanic crust that is subducted under another plate. They are therefore colored an intermediate shade of blue mixed with a pale shade of red (*a*). Oceanic islands associated with rising regions in the mantle have isotopic ratios of both strontium and neodymium that are between those of the upper mantle and those of the continents, and so they are colored an intermediate shade of purple (*b*). They could have been generated from a mixture of the upper mantle and



um in the upper mantle was greater than it was in the crust just after differentiation; the ratio of the isotope neodymium 143 to neodymium 144, however, remained the same throughout the earth. In the diagram half of the samarium 147 decays to neodymium 143 in each part of the earth. Therefore the ratio of neodymium 143 to neodymium 144 grows most rapidly in the upper mantle, less rapidly in the bulk earth and least rapidly in the crust. The darker the red, the smaller the present ratio of neodymium 143 to neodymium 144.



material from the mantle part of the continental lithosphere, which may have sunk into the upper mantle. Oceanic islands could also have been built up from material that leaks from the lower mantle into the upper mantle; they are colored a pale shade of purple, between that of the upper mantle and that of the lower mantle (*c*). The shades of gray of the landforms at the surface have no isotopic significance, and the vertical scale of the schematic diagram has been exaggerated. The motions of the plates are indicated by arrows.

mantle, the depth determined from the age of the plate can be subtracted from the observed depth to obtain what is called the residual depth. It is the long-wavelength variations in residual depth that we hoped could be correlated with the long-wavelength variations in the height of the sea surface.

In the Atlantic, Pacific and Indian oceans a good correspondence has been found between the two kinds of observations [*see illustration on page 71*]. For example, the sea floor around Hawaii is too shallow for its age, and so the residual depth is positive. (Both the estimate of depth from the age of the plate and the measured depth are negative numbers.) The sea surface in the same region is elevated as well, reflecting a positive gravitational anomaly. Both features are predicted to exist above a hot, rising current in the mantle. Similarly, in many regions that are too deep for their age the sea surface is depressed; the mantle under such regions is probably sinking. In all the oceans active volcanic islands are at the center of what are thought to be upwelling regions in the mantle. Perhaps the most striking feature of the observations is the absence of any obvious relation between the motions of the plates and the small-scale circulation of the underlying mantle. Upwelling regions are not in general associated with midocean ridges; they are distributed all over the ocean floor. There are even several areas along ridge axes, such as along the ridge between Australia and Antarctica, where there appears to be a downwelling current in the mantle.

The pattern of rising and sinking regions determined by this method agrees in all essentials with the predictions of the models. The size and spacing of the gravitational anomalies are similar to the calculated ones and, as was predicted, the separation between rising and sinking regions is much less than the size of the plates. The data also show a slight elongation of the pattern of rising and sinking regions in the direction of the plate motion. In other words, the large-scale circulation responsible for the motions of the plates is probably superimposed on the smaller-scale circulation that is drawn out in the direction of the plate motion. The observations therefore support a two-scale model of mantle convection.

Although our investigations of residual depths and gravitational anomalies make it possible to map rising and sinking regions, they provide little information about the depth to which the circulation extends and none at all about its evolution with time. In order to make further progress geochemical observations must be related to the known features of convective motion. Ideally one would like to be able to draw material from the mantle at selected points and match its composition with material elsewhere on the earth. In this way the mixing and distribution of the material could be traced over time. Unfortunately changes in elemental composition that are introduced as the material within the mantle comes to the surface make it difficult to infer the composition of the mantle from the composition of the melt. Progress has come, however, with advances in the design of mass spectrometers that can measure trace amounts of various isotopes.

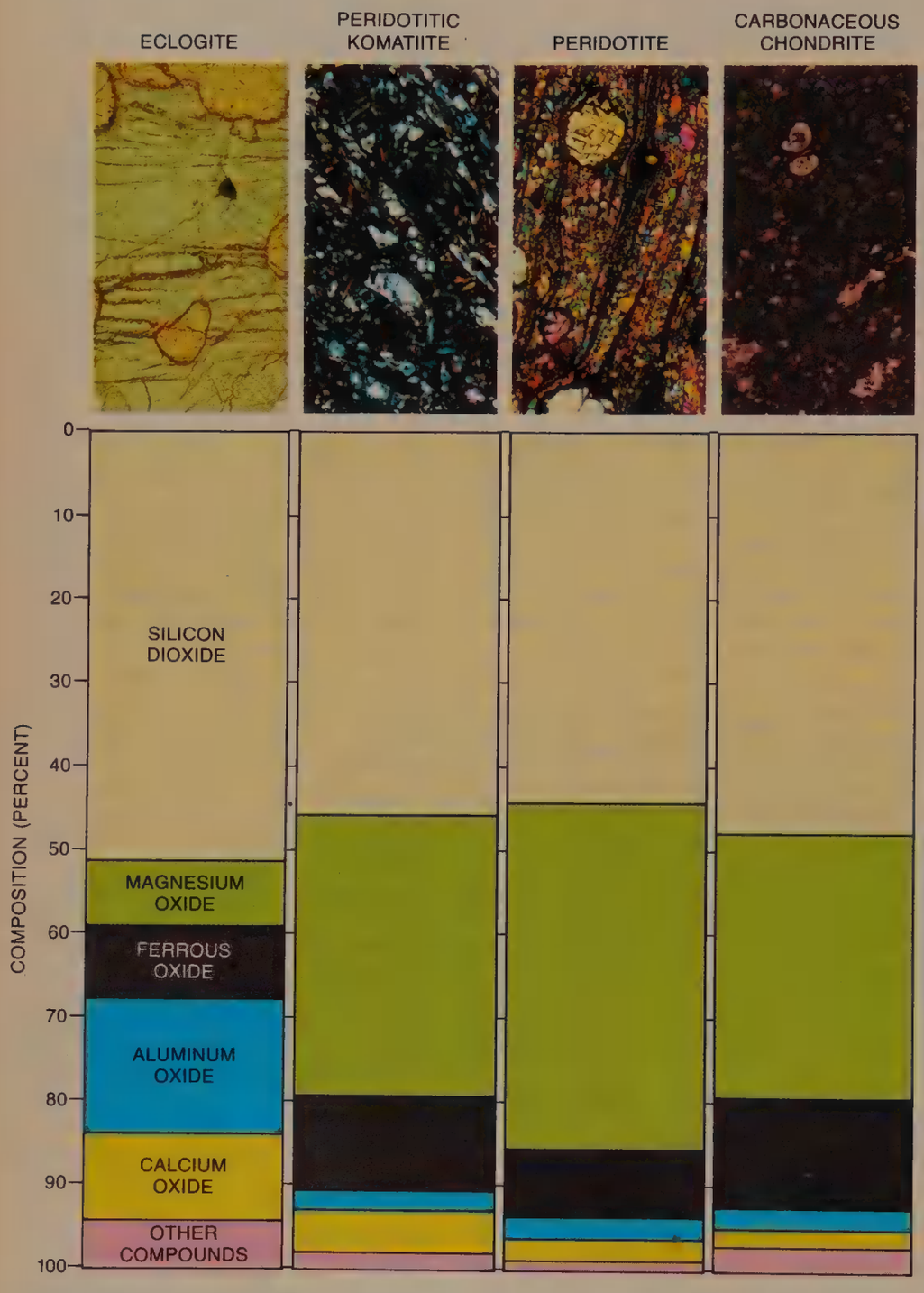
The isotopes of an element are chemically indistinguishable: variation in the number of neutrons in the atomic nucleus has no effect on the electronic structure, which is solely responsible for the chemical behavior of the atom. Hence the ratio of the concentrations of two isotopes of a given element is not affected by the high-temperature chemical reactions that take place in volcanoes. The ratios of the relative abundances of a few isotopes do, however, change with time because certain isotopes are radioactive and decay into isotopes of other elements. Therefore any chemical reactions in the distant past that brought about changes in the relative abundances of different radioactive elements are now reflected in isotope ratios of the decay products.

For example, continental crust was made by the melting of the upper mantle, and the process concentrated in the continents more of the element rubidium than of the element strontium. If for the sake of simplicity it is assumed that the differentiation of continental crust from the mantle took place instantaneously, then just after the melting the ratio of strontium 87 to strontium 86 was the same in the continents as it was in the upper mantle. It happens that rubidium 87 is radioactive; with a half-life of 48 billion years it decays into strontium 87. Therefore as time passes the amount of strontium 87 with respect to strontium 86 increases. It does so both in the upper mantle and in the continents, but the rate of increase is higher in the continents, because after the differentiation of continental crust there is more rubidium 87 in the continents. Continental crust is thereby indelibly marked with a characteristic ratio of strontium 87 to strontium 86 higher than the ratio of the two isotopes in the upper mantle.

Isotope ratios determined by radioactive decay can also be exploited to follow the circulation of material derived from the crust, just as a colored dye added to a mass of water makes it possible to follow its movement underground. For example, it has been found that in island-arc volcanoes a certain fraction of the strontium is from continental crust and that the fraction of neodymium from continental crust in the volcanoes is much smaller. If such volcanoes were directly derived from continental

crust by melting, the isotopic signatures of both strontium and neodymium in the volcanoes would be identical with those in the continents. Since the strontium shows a greater continental influence than the neodymium, there must be some process that transfers strontium from the continents to the source regions of island-arc volcanoes and leaves neodymium behind.

That process begins with continental erosion, which carries most of the strontium in the continental rocks into solution. The isotopic composition of seawater is dominated by that of the rivers, and when seawater circulates through the hot oceanic crust at the axis of midocean ridges, much of the continental strontium is left behind in the basalt of oceanic crust. Most of the neodymium in continental rocks, on the other hand, does not go into solution and remains in sediments. When the oceanic crust is then carried below the island arc by a sinking plate, it melts to yield a rock enriched in continental strontium but not in continental neodymium.



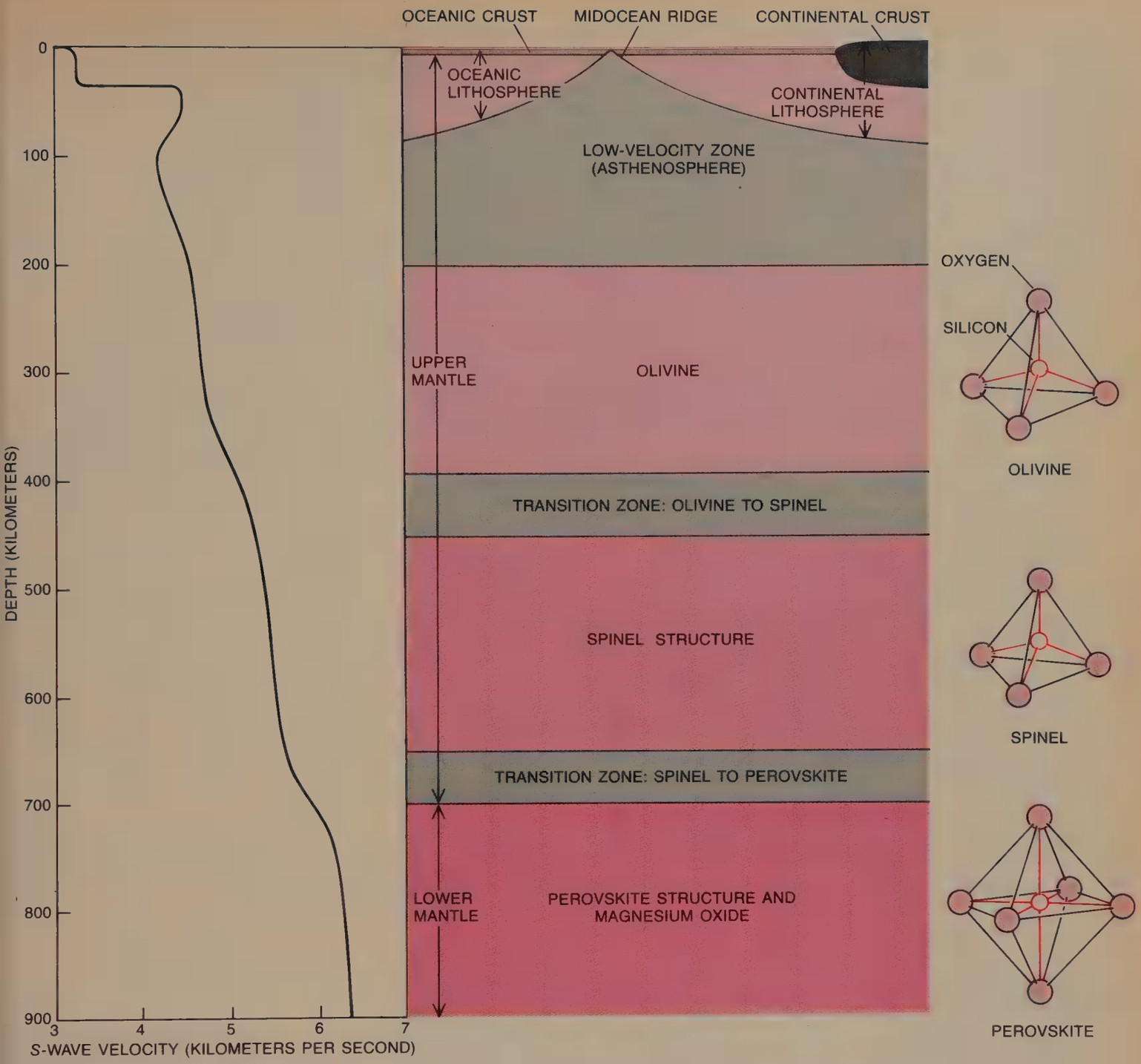
ROCK FROM WHICH THE MANTLE IS MADE must be capable of producing the melts that are known to come from it. The basalt now being erupted along midocean ridges could be formed by the complete melting of an eclogite, whose composition is shown below the photograph at the left. In the first half of the earth's history, however, melts much richer in magnesium were erupted, which cooled to form peridotitic komatiite. The long, thin crystals in this section show that this rock was once entirely molten. Such melts could have come from the rock peridotite, whose composition is much closer to that of a peridotitic komatiite than the composition of eclogite is. The composition of peridotite also resembles the oxide composition of carbonaceous chondrites, extremely old meteorites that are thought to be similar in composition to the material from which the earth was formed. Carbonaceous chondrites contain water, carbon, metallic iron and nickel and sulfides as well as oxides, but these materials have not been included in the comparison because they were largely separated from the mantle in the early differentiation of the earth. The photomicrographs of eclogite and peridotite were provided by F. R. Boyd of the Carnegie Institution of Washington, the photomicrograph of peridotitic komatiite by Euan G. Nisbet of the University of Saskatchewan and that of carbonaceous chondrite by Lawrence Grossman of the University of Chicago. The compositions given for comparison are typical, but they may differ slightly from those of the specimens illustrated.

There is a striking uniformity among all measurements of isotopic ratios made on basaltic rocks from midocean ridges, all of which show the depletion caused by the formation of continents. Such uniformity calls for vigorous mixing, and so it is natural to ask whether thermal convection could be responsible. Nicholas Hoffman of the University of Cambridge has used numerical models of mantle convection to follow the evolution of a square sample of the mantle fluid as it is deformed. He found that the mixing is vigorous, and that it can easily produce the observed uniformities in isotopic ratios within a billion years.

It is less straightforward to trace the source of the volcanic material erupted above the hot, upwelling regions of the mantle. For example, the isotopic ratios of volcanic material from Hawaii are intermediate between those found along midocean ridges and those thought to be characteristic of the earth as a whole, whereas those from Kerguelen in the southern Indian Ocean must come from a region enriched in both rubidium and samarium with respect to the bulk earth. To account for such anomalous ratios the material must have been isolated from the convecting part of the mantle for about a billion years. There are two regions that could give rise to the observed ratios: continental lithosphere and any region of the mantle, such as its lower part, that does not take part in upper-mantle convection. It is not yet clear whether either source is consistent with all the observations.

There are now several lines of evidence suggesting all scales of convection that have been observed in the upper mantle are confined to a layer whose lower boundary is no deeper than 700 kilometers. Bear in mind that the numerical simulation, which successfully predicts the distribution of rising and sinking regions for the small-scale convection cells, assumes the material circulates to a depth of no more than 700 kilometers. Isotopic ratios can also yield a rough estimate of this depth. The argument is ingenious but somewhat complicated, and it depends on knowing the total amount of the samarium and neodymium in the crust as well as in the earth as a whole.

The samarium and neodymium in the



SEISMIC PROFILE OF THE UPPER MANTLE shows a series of zones in which the wave velocity is relatively constant, alternating with transition zones in which the velocity increases with depth. Laboratory simulations of the pressures within the upper mantle suggest the zones are the result of changes in the solid phase, or rearrangements of the atoms that make up the crystalline structure of the solid, rather than a result of changes in composition. The material between the base of the crust and a depth of about 390 kilometers is peridotite, whose dominant mineral is olivine. Each silicon atom in the olivine structure is surrounded by four oxygen atoms (*top right*). Below 390 kilometers is a transition zone in which the atoms in olivine are rearranged into a denser structure resembling that of the mineral spinel. Each silicon atom in the spinel structure is also surrounded by four oxygen atoms (*middle right*). The deepest major phase change is found at pressures corresponding to a depth of 700 kilometers; the spinel structure breaks down into a mixture of an even denser structure resembling that of the mineral perovskite (*bottom right*) and magnesium oxide (not shown). Each silicon atom in the perovskite structure is surrounded by six oxygen atoms. The low-velocity zone near the top of the mantle may contain a small quantity of melt.

crust can be determined from the many detailed measurements that have been made on a variety of crustal rocks. The composition of the bulk earth can be estimated from the composition of the type of meteorites called carbonaceous chondrites, which are older than the earth. The earth presumably formed out of a cloud of gas and dust or out of meteoritic bodies with a composition similar to that of carbonaceous chondrites. Although the process of formation is not well understood, it is likely that it thoroughly mixed the elemental raw materials. Therefore when the crust differentiated from the mantle, any concentration of an element in the crust would have resulted in a corresponding depletion of the element in the mantle. Since the isotopic ratios in the mantle can be measured, it is possible to calculate the volume of the mantle that had to be depleted in order to produce the observed quantity of the element in the crust.

To understand how the calculation is carried out consider two extreme cases. In the first case the isotopic ratio of neodymium 143 to neodymium 144 in the mantle is found to be infinite; in other words, there is no neodymium 144 in that part of the mantle from which the crust was produced. Such a ratio could come about only if all the neodymium in the convectively mixed part of the mantle had been removed in the course of the differentiation of the crust from the mantle. Neodymium 143 would then be generated within the depleted mantle

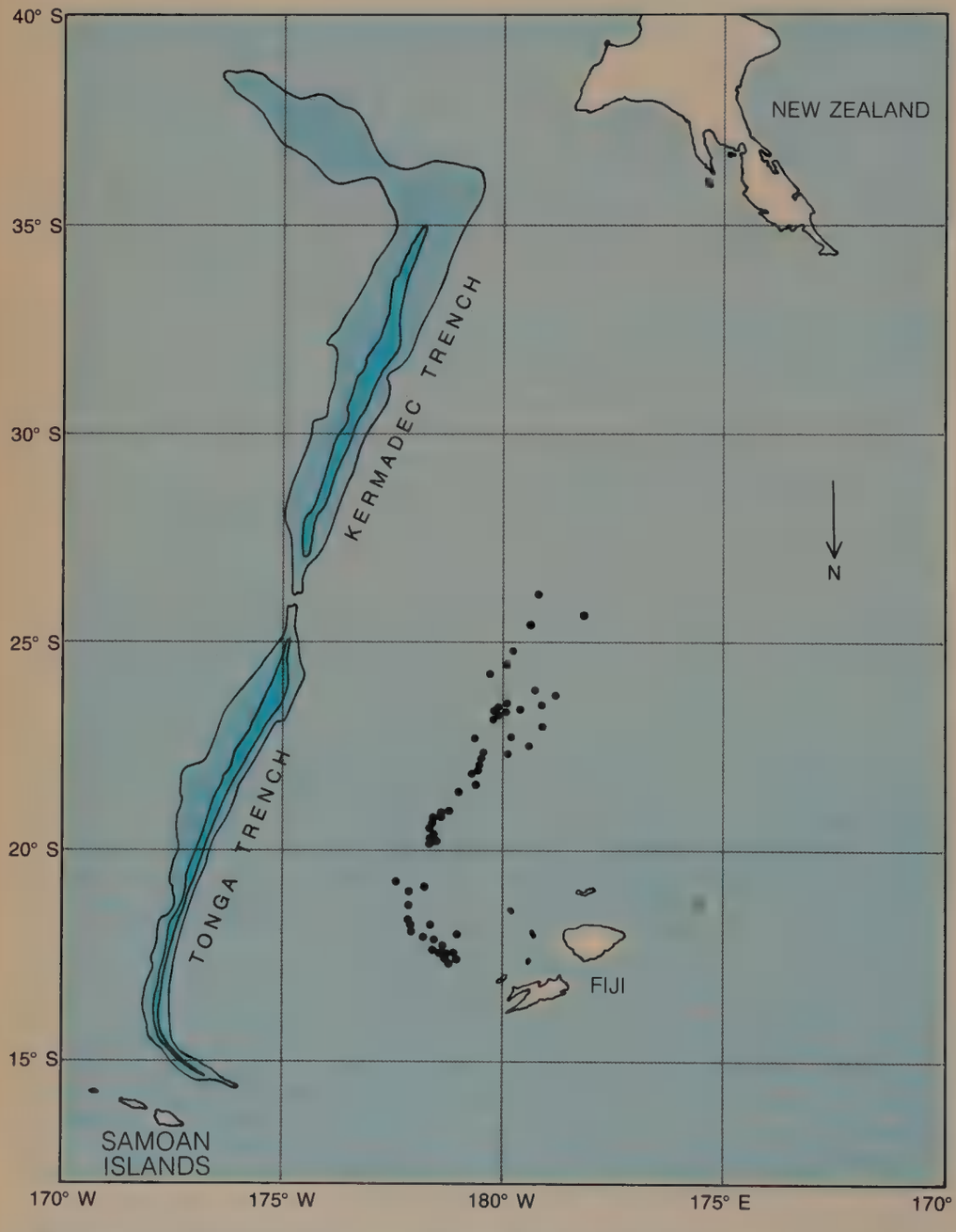
only through the decay of the isotope samarium 147. The volume of mantle involved in crustal production would therefore have to be just enough to provide the quantity of neodymium present in the crust. Since the concentration of neodymium in carbonaceous chondrites can be measured and the total amount of neodymium in the crust can be estimated, the mass of mantle involved is easily obtained.

In the second case the ratio of neodymium 143 to neodymium 144 in the mantle is found to be nearly the same as that in the bulk earth, even though the enrichment of neodymium in the crust is the same as it was in the first case. Here the amount of neodymium extracted from the mantle by the formation of the crust must have been negligible compared with the amount left behind in the mantle. Therefore the volume of mantle involved must have been relatively large.

The real situation lies between the two extremes. The observed isotopic ratios show that the crust has been extracted from about a third of the mass of the mantle. In other words, only a third of the material in the mantle, thoroughly mixed, is needed to account for the relative abundances of the elements and their isotopes in the continents. Since a third of the mantle lies between the base of the crust and a depth of 700 kilometers, the composition of the crust and the mantle is consistent with the idea that convection in the upper 700 kilometers of the mantle is physically separated from any convection in the lower mantle. Moreover, because much of the continental crust has existed for at least two billion years, there must have been only limited transport between the upper and the lower mantle in that period.

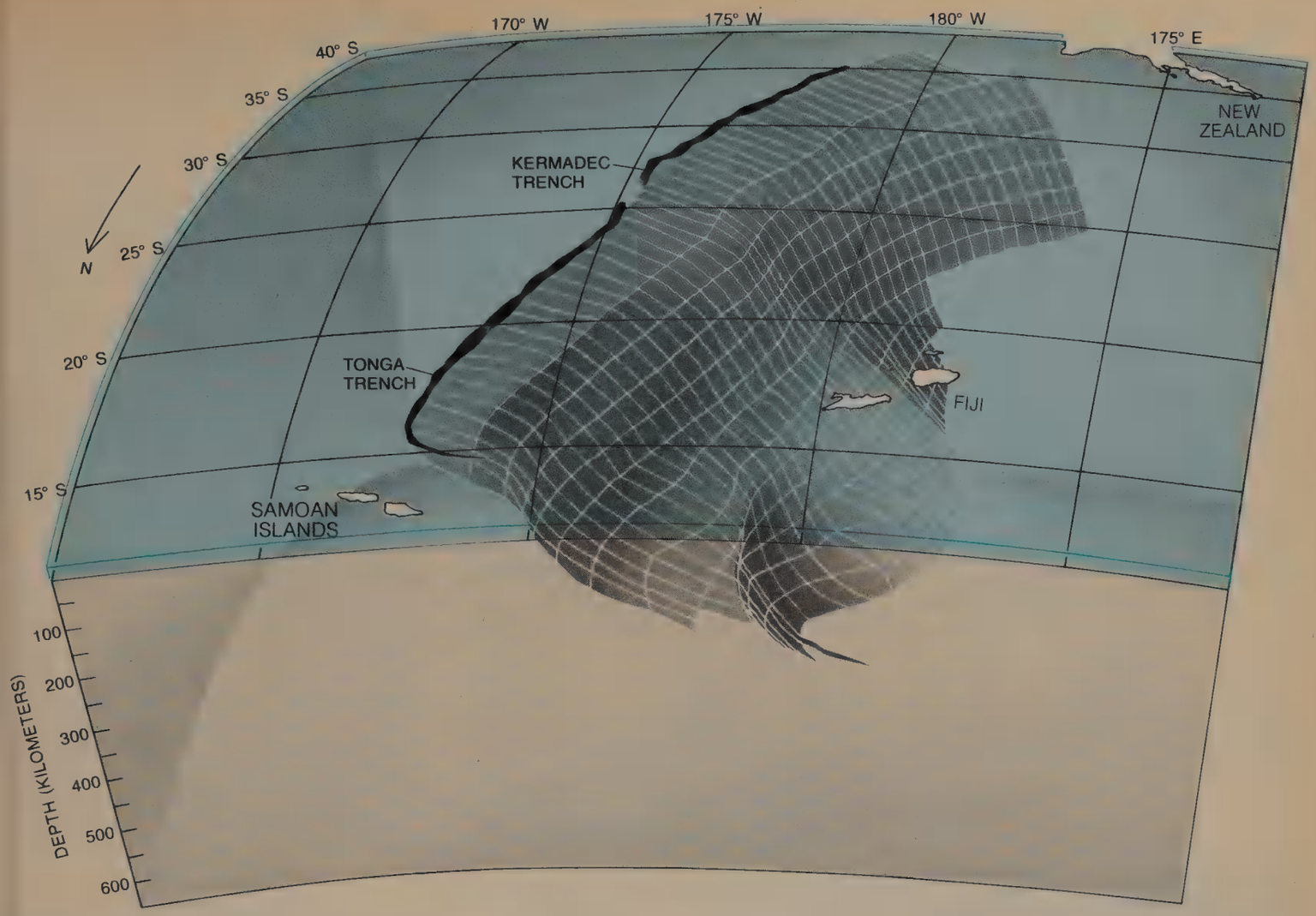
In order to understand why the upper mantle may convect as a separate layer one must consider its composition. Whatever the mantle is made of must be capable of melting in whole or in part to produce basalt, the commonest type of volcanic rock. The entire ocean floor is covered with a layer of basalt at least three kilometers thick, which is generated by melting of the mantle under the midocean ridges. Huge volumes of basalt are also erupted by volcanoes far from the ridges. The most straightforward way to imagine making a basalt is to start with a solid rock of the same composition and melt it entirely. An appropriate rock is eclogite, which consists of garnet and pyroxene, dense minerals that are stable at high pressures.

A number of investigators still believe large parts of the mantle are made of eclogite, but there are serious difficulties with this view. In the first half of the earth's history much of the volcanic material erupted at the surface was too rich in magnesium to have been produced by the melting of an eclogite. Such a melt, however, could readily have come from peridotite, a rock whose dominant mineral is olivine, a magnesium silicate (Mg_2SiO_4). Fragments of peridotite are carried from great depths to the surface by basaltic magma, and they also dominate the rocks found in diamond-bearing kimberlite "pipes," volcanic structures that must originate in the mantle at depths of at least 150 kilometers. Most important, the composition of the carbonaceous chondrites agrees well with the composition of peridotite but not with that of eclogite.



EARTHQUAKE EPICENTERS, or surface coordinates, are plotted to the west of the Tonga and Kermadec trenches in the South Pacific for all earthquakes detected at depths of from 500 to 600 kilometers below the surface. The deep earthquake foci outline a shape similar to the shape of the Tonga trench. The resemblance shows that the material at the base of the trench and the material at depth are closely connected. The diagram is based on a chart prepared by by Lynn R. Sykes of the Lamont-Doherty Geological Observatory of Columbia University.

If the composition of the mantle resembles that of peridotite, laboratory tests show there must be two changes of solid phase in the rock that forms the upper 700 kilometers of the mantle. A phase change in a solid involves a reordering of the atoms of the crystal structure in response to a change in pressure or temperature. Experiments show that the crystal structure of olivine changes to that of a spinel at pressures corresponding to depths between 390 and 450 kilometers. This change increases the density of the material by about 10 percent. The second phase change converts the spinel structure into that of a perovskite, and the density of the material increases by another 10 percent. The differences between the perovskite structure and the spinel are considerably more important than the differences between the spinel structure and the olivine. The silicon atoms in the perovskite structure are each surrounded by six oxygen atoms, whereas the silicon atoms in both the olivine structure and the



FOCI OF EARTHQUAKES at all depths to the west of the Tonga and Kermadec trenches are fitted by a computer to a smoothly descending surface. The surface can be understood as being the central plane of a descending slab of the lithosphere, which is subducted into the mantle under an island arc. The general shape of the island arc is retained in the slab at all depths. It is now thought the island arc overrides the sinking slab and shapes it like the cutting tool of a lathe. The existence of such slabs is strong evidence for relatively rapid circulation within the mantle. The diagram is based on a computer-generated surface constructed by Bryan L. Isacks of Cornell University.

spinel are each surrounded by only four oxygen atoms. The phase change to the perovskite structure takes place at the pressure that corresponds to a depth of about 700 kilometers.

The two phase changes coincide with transition zones in the mantle observed by seismologists. The velocity of seismic waves generated by earthquakes increases sharply between 390 and 450 kilometers and again near 700 kilometers. It is also striking that earthquakes have not been found deeper than 700 kilometers. Remember that such earthquakes mark the greatest depth to which a subducted slab of cold mantle can descend without becoming assimilated into the surrounding material.

The progress in the understanding of the dynamics of the mantle has encouraged a number of geophysicists to study the thermal history of the earth. Thermal evolution is regulated by the mantle because the predominantly iron core is an excellent conductor, whereas the mantle is not. In the mantle heat is transported primarily by convection. The largest source of heat loss is the process of plate creation at midocean ridges. The hot plate is cooled partly by the seawater circulating through the rock near the ridge axes and partly by thermal conduction through the sediments that cover the older parts of the plate. The total rate of heat loss is about 40 million megawatts.

Heat is generated in the earth's interior by the decay of the radioactive isotopes of uranium, thorium and potassium. The rate of such heat generation from uranium and thorium can be estimated with some confidence from the composition of carbonaceous chondrites. Potassium, however, is a volatile element, and it is less abundant in the earth than it is in carbonaceous chondrites. Its terrestrial abundance must be estimated from the composition of crustal rocks. The total heat-generation rate estimated in this way is about 20 million megawatts. Hence the earth must be cooling. There is even some direct evidence that the mantle has cooled with time. The volcanic rocks whose high magnesium content excluded eclogite as a source rock were erupted in quantity only during the first half of the earth's history. Because of their composition, at the time they were extruded they must have been about 200 degrees Celsius hotter than any volcanic rocks now being erupted.

The heat that is being lost comes partly from higher rates of radioactive heat generation in the past and probably also partly from heat generated by the formation of the earth. If the earth were a perfect conductor of heat, the rate of heat loss would always have exactly matched the rate of heat generation. The fact is, however, that the mantle is a poor conductor and can store heat, which is slowly released as it and the core cool down. The response time of the earth's heat loss to changes in its heat generation is a measure of the mantle's overall insulating properties.

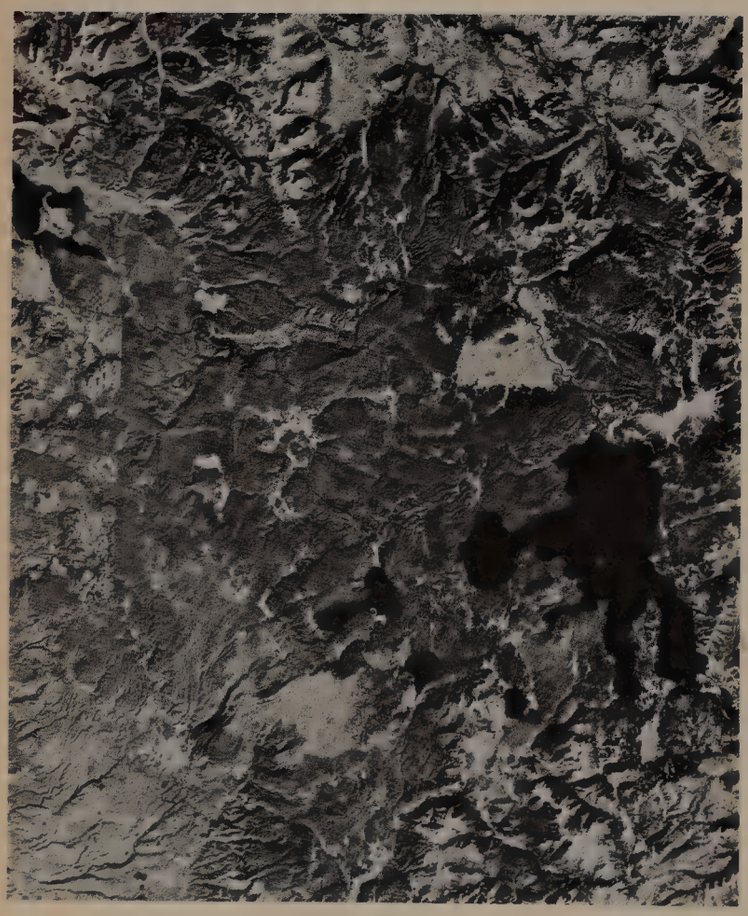
Calculations have now been carried out on the assumption that convection currents carry heat throughout the mantle, and they give a response time that is too short. In other words, convection in the entire mantle would carry heat to the surface too efficiently to be consistent with the observed ratio of heat loss to heat generation. In this model the rate of

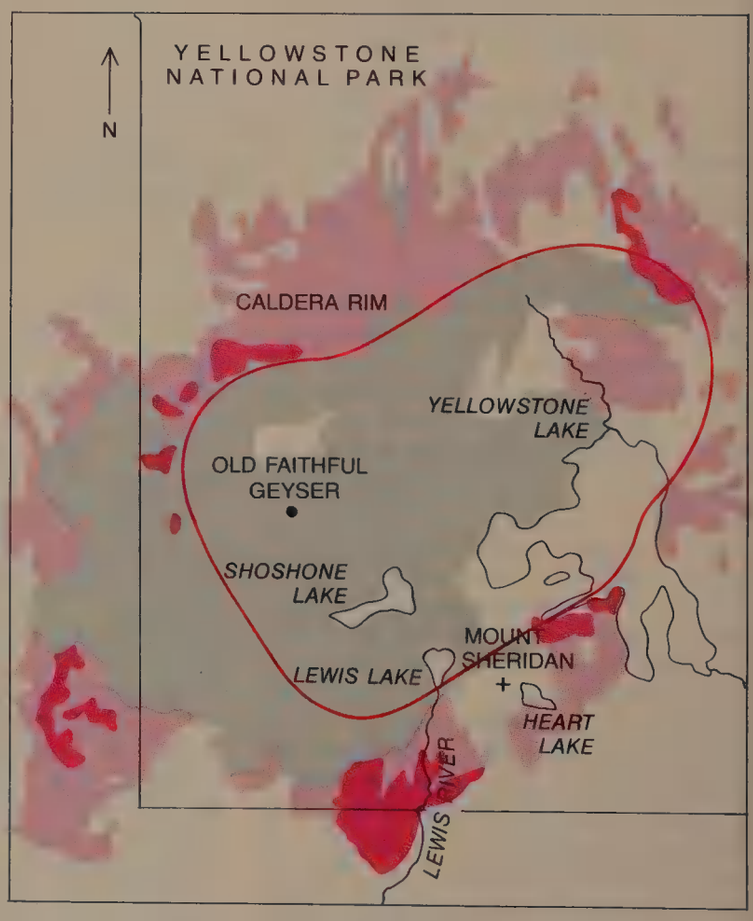
heat loss could never be double the rate of heat generation because heat could not be stored in the earth for a sufficiently long time. On the other hand, if the circulation in the upper mantle is separate from the circulation in the lower mantle, the upper mantle can act as a thermal insulator. The heat stored in the lower mantle can thereby be retained. In principle the insulating mechanism is similar to domestic double glazing: the rate of heat loss from a house is decreased by introducing a convective layer of air between two panes of glass.

There is by no means unanimous agreement among geophysicists regarding the preceding conclusions or any other conclusions that depend on the properties of the lower mantle. Indeed, so little is known about the mantle at depths below 700 kilometers that there can be few constraints placed on theoretical speculation. The composition of the lower mantle is thought to be similar to that of the carbonaceous chondrites, but it is not yet possible to exclude increases in the ratio of iron to magnesium of as much as 5 percent. Such a change would increase the density by about 2 percent, in addition to the increase in density caused by the phase change from the spinel structure to the perovskite one. Even such small changes in density could have a profound influence on the interchange of material between the upper and the lower mantle. Within the lower mantle there is no evidence for any major phase changes, although it would not be surprising if some minor transition zones have so far escaped detection. Within about 200 kilometers of the core, however, the velocity of seismic waves is variable, and the waves are scattered in all directions. The cause of the variations is not clear, and no systematic attempt has yet been made to map the velocity changes and relate them to other features.

Recently a simple but ingenious device called the diamond-anvil cell has become available, which makes it possible for the first time to subject materials in the laboratory to the enormous pressures encountered in the lower mantle and in the core. A small sample of a mineral is placed between the two sharp points of gem-quality diamonds, and the material is squeezed between the points by turning a screw by hand. Although the force that generates the pressure on the sample can therefore be quite small, the area across which the force is applied is so small that the pressure between the points of the anvil can be far greater than the pressure generated by a large hydraulic press. By analyzing minute samples compressed by the anvil new phase changes have been detected that may occur in the lower mantle.

Twenty years ago many earth scientists considering the evidence for or against continental drift were consciously or unconsciously thinking in terms of a static model of the earth. This situation changed completely with the general acceptance of sea-floor spreading and plate tectonics. The effect on the study of the dynamics of the mantle was particularly profound, since plate tectonics established the existence of mantle convection without providing much information about the forces involved. Some of the first attempts to understand mantle dynamics limited the circulation to the plate motions and a return flow that carried the mantle material from trenches to ridges. The dynamic models and observations of the gravity field have now clearly shown that much of the convective circulation is not related to the movements or boundaries of the plates. I believe we now understand the outlines of the dynamics of the upper mantle; the challenge is to discover how the more massive lower mantle behaves.





YELLOWSTONE CALDERA is one of the few volcanic structures in a continental region known to be associated with a long-wavelength gravitational anomaly and therefore with a hot, upwelling region in the mantle. In most continental areas convection in the mantle has not been mapped because the gravitational field is too poorly known. A pair of special satellites orbiting the earth at a height of less than 200 kilometers could make a worldwide map of the long-wavelength gravitational field; volcanoes that are directly related to motions in the mantle could thereby be distinguished from those that are not. Such a project is now being considered by NASA. The dark-colored areas in the map at the right are accumulations of rhyolite (a glassy volcanic rock) that began to erupt 1.2 million years ago, before the region inside the rim of the caldera collapsed; the light-colored areas are rhyolitic ash flows deposited during the climactic collapse of the roof of the magma chamber under the caldera. The gray areas mark subsequent eruptions. The satellite image was made by Landsat.

We gave the instant print a split personality.

Now there are two ways to enjoy instant pictures.

Kodak scientists have learned to manage a phenomenon called "controlled adhesion," the stripping-layer concept used for years as a laboratory tool to study how color image formation occurred in different layers of film. And we've turned the concept into a new product with broad consumer appeal.

The product, Kodamatic™ Trimprint™ instant color film, has an integrated stripping agent which allows picture-takers to separate the fully developed print from the backing like a pressure-sensitive label. With backing removed (and imaging elements discarded), the print is slimmer, 50% lighter, and can easily be cut, posted, thumbtacked, mailed, carried or displayed in an album or photo cube.

Development of a film-stripping layer that was reproducible and controllable resulted from the combined efforts of Kodak chemists and engineers, as well as manufacturing and marketing specialists who collaborated in adapting technology to meet market demands.

Our challenge was threefold: First, to find a permeable adhesive material (strong enough to bond the picture unit, yet flexible enough to permit quick and easy separation) that could be sandwiched between the image-forming and image-receiving layers. Second, to design machinery to slit (with finite accuracy) the separation point on the specification layer of the instant print. Third, to add this feature, yet retain all other attributes of the film, regardless of whether the prints would be separated or left as integral units.

While our dry-strippable capability may be beneficial to you today only as a consumer, tomorrow it may prove useful in industrial or scientific applications.

Kodak

Kodak. Where technology anticipates need.



© Eastman Kodak Company, 1983

CAN YOU FIND RISKS IN TH

No matter what business you're in, from banking to satellite communications, from petroleum to wind farms, from medicine to robotics, the computer chip is unavoidably a basic part of how you do business. And if you do business.

In the greatly magnified chip shown here, error is easy to spot. Potential risk is not.

How, for example, does an insurance company identify and analyze the infinite risks associated with technological developments like this over the course of years to come?

How do you insure the billions of microscopic chips that help fly airplanes, power pacemakers, manage assembly lines, transfer bank deposits and perform a thousand other vital functions every day?

At AIG it is our business to understand the miracles as well as the dangers inherent in trusting the nation's economy to the ubiquitous microchip.

That's why our engineers and underwriters are among the industry's most experienced in evaluating the newest and most complex challenges High Tech companies can offer.

After all, if business is going to move forward, we have to be ready to underwrite the necessary risks. The High Tech revolution won't stop just because we don't know what dangers it may bring.

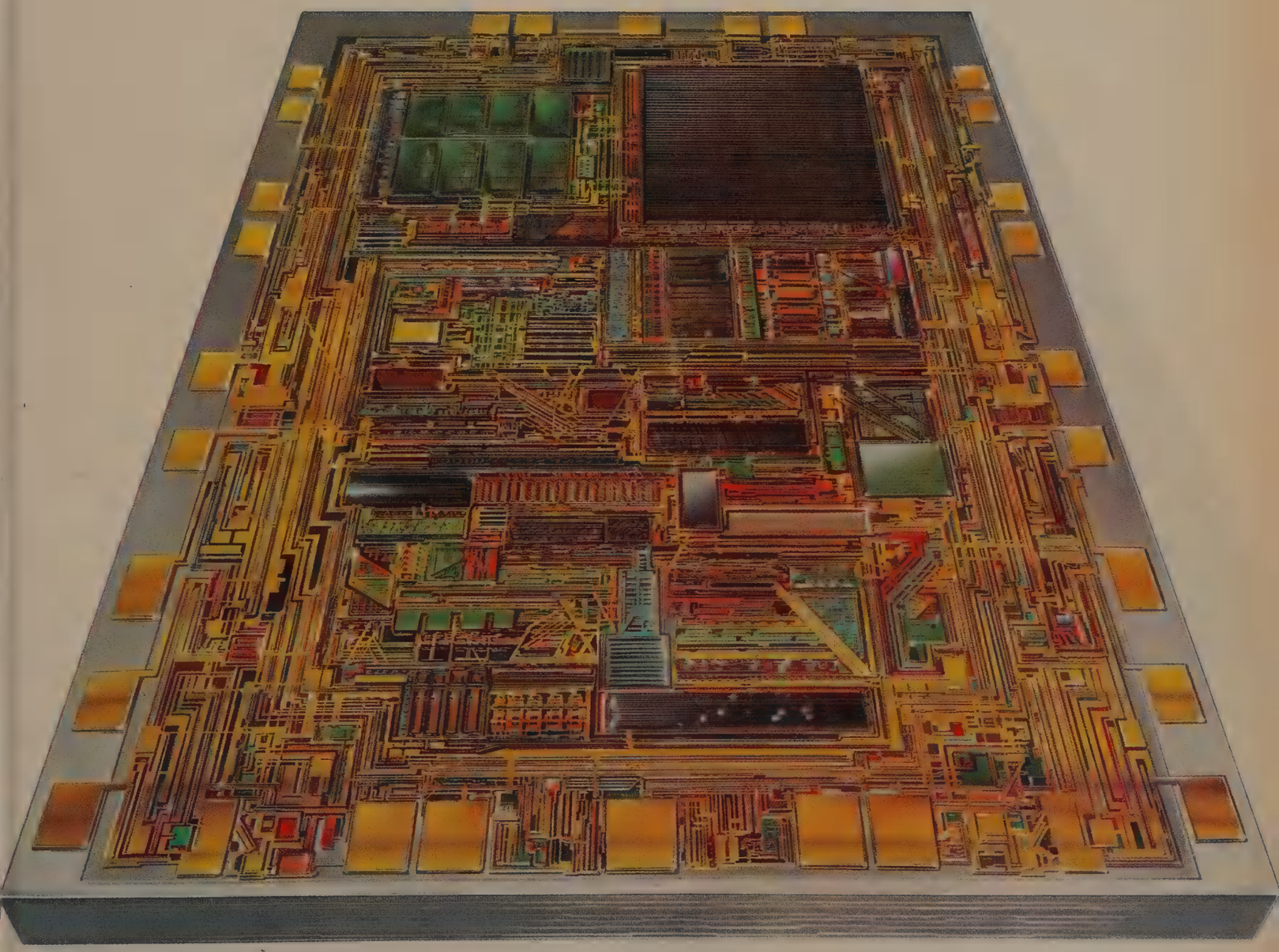
For over 60 years the AIG Companies have been helping to protect investments which have built the future. We know that if business is to take advantage of what science discovers, we have to be ready to assume the necessary risks.

Our experience in over a thousand different fields has made us one of the most successful and innovative international insurance organizations in the world. Today, this not only means looking to the future, but knowing how to assure your place in that future.

ONCE AGAIN THE ANSWER IS AIG.

THE AIG COMPANIES: *American International Underwriters, American Home Assurance Co., Commerce and Industry Insurance Co., National Union Fire Insurance Company of Pittsburgh, Pa., American International Life Assurance Co., North American Managers, New Hampshire Insurance Co. and over 230 other companies operating around the world to meet your insurance needs.*

THE HIGH TECH IS PICTURE?



SCIENCE AND THE CITIZEN

Science or R&D

What does the proposed Federal budget for 1984 reveal about the current state of the support of science in the U.S.? It all depends on how science is defined. If "science" includes the development, testing and evaluation of new technologies for national defense and industry, then science is doing well: Federal outlays for research and development (R&D) should increase 13.3 percent in constant (inflation-corrected) dollars, even in the teeth of another projected $200 billion Federal deficit. If science is defined as the search for new knowledge, then the picture begins to blur. Since the Reagan Administration began reshuffling the R&D budget in 1981, funding for defense R&D has increased 82 percent, so that it now accounts for more than 65 percent of the total Federal R&D budget. The dollar amount for nondefense R&D has remained virtually constant since it was cut back $3 billion in 1981, and so the investment in nondefense science has actually been diminished by inflation.

The President's science adviser, George A. Keyworth II, has his own views on how support should be given. He has suggested that the scientific community choose the priorities, both within disciplines and among them. The first it has begun to do, notably through the Committee on Science, Engineering, and Public Policy (COSEPUP) of the National Academy of Sciences; the second it has not done. Keyworth has said the nondefense areas that deserve high priority are the physical sciences, mathematics and engineering. In this view the salvation of the U.S. economy, its position in world markets, its defense and science itself lies in high technology.

The budget category that best serves as an indicator of Federal support in the core sciences (the ones that are taught in universities) is "basic research." Next year the biggest boost to basic research will come from the National Science Foundation, which traditionally supports all sciences except the life and clinical sciences. The NSF is scheduled to receive a 17.4 percent increase in its overall budget and so to be restored almost to its 1981 funding level. Even so, some NSF programs, such as International Cooperative Scientific Activities, will receive large cuts in order to allow substantial increases for engineering, technology development and certain scientific areas: "information sciences," "astronomy and atmospheric sciences" and deep-sea geologic drilling (up 60 percent).

The Administration apparently sees the NSF as a vessel for its new policy on science education. For some years universities have been calling attention to the physical deterioration of their research facilities. Moreover, there has been a steady decline in Federal support of graduate education and in the training and support of young faculty members. The Administration budget calls for the following NSF correctives: a 61 percent increase in grants for instrumentation, a 20 percent increase in grants for graduate research fellowships, Presidential Young Investigator Awards as an incentive for young faculty members (the funds for which are to be matched by industry) and a variety of programs to provide an incentive for performance in precollege teaching, particularly in mathematics and engineering.

R&D expenditures by the Department of Defense have traditionally included some support for "basic research." The Defense Department R&D budget is scheduled to increase 30 percent in 1984, to $33 billion. Basic research in the department is funded through the "6.1 category," which is scheduled to rise 9.2 percent in current dollars. An increasing fraction of that money, however, is going to applied research. Moreover, sentiment in Congress is said to be running against a 30 percent increase in funds for Defense Department R&D. It is more than likely the House and Senate appropriations committees will, as they did last year, make large cuts in the defense budget. Defense Department budget makers may have learned to expect such cuts, because the requested increase for R&D is $2.2 billion more than the department's predictions made last year, and $400 million more than the cut made by Congress in the 1983 R&D budget.

The Office of Science and Technology Policy, headed by Keyworth, and the Office of Management and Budget, headed by David A. Stockman, have been urged by the White House to reduce funding, at least for nondefense R&D, of short-term projects that are more appropriately supported by industry. Therefore the Department of Energy (while it still exists as a separate agency) will seek less money for research in conventional energy supplies. It will seek more money, however, for research in nuclear energy and high-energy physics. The budget for the National Aeronautics and Space Administration will decrease 3.1 percent (including the effect of inflation), although there are increases scheduled for some important scientific projects, such as the Space Telescope, the Gamma Ray Observatory and the new Venus Radar Mapper. On the other hand, NASA's development of scientific satellites to make observations of the land, the sea and the atmosphere will virtually come to a halt.

What about the life sciences at a time when the emphasis is intended to be on the physical sciences, mathematics and engineering? According to Keyworth, the life sciences have been receiving "rather healthy funding." The proposed 1984 budget for the National Institutes of Health, however, shows a constant-dollar decrease of 3.8 percent. The NIH supports more than half of all the basic research done at universities in the U.S. Over the past 10 years, which have seen an explosive growth in molecular biology, the NIH budget has increased a total of only 5 percent, and it has declined steadily since its high point in 1979.

The prime directive at the NIH is the funding of investigator-initiated competitive research grants, awarded under a peer-review system. During the Carter Administration the NIH was expected to make a minimum of 5,000 new research grants per year, regardless of the funding level. That goal has never been reached. With the arrival of a new Secretary of Health and Human Services, Margaret Heckler, the Administration has submitted a budget it hopes will enable the agency to reach the 5,000-grant level. In order for the NIH to do so without exceeding its budget, however, it will have to cut other programs; for example, training grants (primarily postdoctoral) and support for research centers across the country will drop more than 10 percent. By the same token the average size of a research grant will have to be decreased 10 percent. Moreover, the NIH has begun to deny the universities reimbursement for indirect costs: administrative expenses associated with research programs. The NIH maintains that these expenditures have gone up unreasonably over the past few years, and that the universities must pay a larger fraction of the costs.

Overall Federal support for research at universities has dropped 3.3 percent since 1981. The budget indicates that even though the NSF may get a 17.8 percent increase, the funding of university-based research by Federal agencies will continue to decline next year.

The Administration budget may not be the final word on the support of basic research in 1984. The Association of American Universities, together with 120 other research organizations, has submitted a proposal that would increase the NIH budget by $414 million. The two Missouri senators, John C. Danforth (Republican) and Thomas F. Eagleton (Democrat), have submitted a bill, "The University Research Capacity Restoration Act of 1983," that

DISCOVER THE TASTE OF THE RENAISSANCE.



·AMARETTO DI SARONNO·ORIGINALE·1525·

FOR FREE FOOD AND DRINK RECIPE BOOKS, WRITE: BOOKLETS, P.O. BOX 99085, LOUISVILLE, KY. 40299.

LIQUEUR ... PROOF. SOLE U.S. DISTRIBUTOR, FOREIGN VINTAGES, INC., N.Y., N.Y. © 1982 / A UNIQUE PRODUCT FROM ILLVA SARONNO, ITALY

NUCLEAR ELECTRICITY
Who Stands Where



At a time of intense criticism of nuclear-generated electricity, an independent academic study found surprisingly strong support among persons considered well informed on the subject.

The study was part of a large project directed by Professors Stanley Rothman and S. Robert Lichter and sponsored by the Research Institute of International Change at Columbia University, Smith College, and the Graduate Program in Science, Technology, and Public Policy at George Washington University.

A survey of scientists

In mid-1980, little more than a year after the accident at Three Mile Island, a detailed questionnaire on energy issues was sent to a purely random sample of: (1) scientists listed in *American Men and Women of Science,* (2) scientists in energy-related disciplines, and (3) scientists in fields closely related to nuclear energy.

Asked how best to proceed with nuclear energy (see Table 1), an overwhelming majority from all groups felt that the nation should continue with the development of nuclear energy.

Table 1
VIEWS OF SCIENTISTS ON HOW TO PROCEED WITH NUCLEAR ENERGY

Random Sample	Energy Experts	Nuclear Experts	
53%	70%	92%	Proceed rapidly
36	25	8	Proceed slowly
7	4	0	Halt development
3	1	0	Dismantle plants

Most felt we should proceed rapidly.

This high level of support is echoed in Table 2, which notes positive attitudes toward the acceptability of risk posed by nuclear plants; the scientists' personal willingness to have a nuclear plant located where they live; and the scientists' confidence that enough knowledge exists to solve the scien-

Table 2
SCIENTISTS' ATTITUDES ON NUCLEAR ENERGY ISSUES

Random Sample	Energy Experts	Nuclear Experts	
65%	75%	99%	Risks acceptable
69	80	98	Willing to locate nuclear plants in their cities
75	91	100	Enough knowledge to solve nuclear problems

tific and technical problems posed by nuclear energy.

Expanding the study

Such positive results were somewhat surprising to the researchers. They decided to expand their study. This was a time when public concern was high and the nuclear regulatory environment was toughening considerably, so the researchers wondered whether decision makers in the nuclear field would be as wary of nuclear power plants as the man in the street.

To find out, the researchers mailed questionnaires to top decision makers in seven different categories:

- the nuclear power industry
- the financial community
- the Nuclear Regulatory Commission
- other regulatory agencies involved with nuclear energy
- members of Congress who were involved with nuclear policy
- outside experts (scientists, social scientists and consultants involved with nuclear energy issues)
- groups with professed anti-nuclear views.

The results were again surprising. As Table 3 indicates, outside experts and financiers were as united in their support of nuclear energy development as were industry executives. The only significant opposition came from the heads of the anti-nuclear groups. A few members of Congress and some regulators also expressed opposition. (It should be noted that response from Congress was low in number.)

Table 3
POLICY PREFERENCES ON NUCLEAR DEVELOPMENT

Anti-Nuclear Groups	Industry	Financiers	NRC	Other Regulators	Congress	Outside Experts	
0%	93%	94%	65%	54%	40%	69%	Proceed rapidly
0	5	6	35	25	40	26	Proceed slowly
33	2	0	0	8	20	5	Halt development
67	0	0	0	13	0	0	Dismantle plants

Table 4 reinforces the pattern. Majorities of all decision-making sectors except the anti-nuclear groups believed nuclear plants safe, the risks acceptable, the problems solvable, and they would be willing to live near a reactor.

The questionnaire also sought to learn what potential problems within the industry were considered most serious. Most of the anti-nuclear groups' leaders rated most of the problems as "very serious." None of the problems was considered that serious by a majority of the other decision makers. Only high-level waste disposal was considered very serious by a majority of persons within any of these seven groups.

The Nuclear Waste Policy Act of 1982 has addressed this problem by establishing a procedure and a timetable leading to the safe disposal of nuclear waste. The concept of burying the waste in stable geological formations deep in the earth has been endorsed by the National Academy of Sciences.

Table 4
GENERAL ATTITUDES TOWARD NUCLEAR ENERGY

Anti-Nuclear Groups	Industry	Financiers	NRC	Other Regulators	Congress	Outside Experts	
100%	2%	0%	4%	37%	25%	12%	Risks unacceptable
0	94	61	72	60	43	74	Very confident we can solve problems
100	2	0	7	28	28	5	Plants unsafe
0	97	83	94	60	53	88	Would live near reactors

What sources can we count on today?

There was more agreement when the leaders were asked which energy sources would make the greatest contributions to our needs by the year 2000. The questionnaire listed 16 possible sources ranging alphabetically from biomass to wind power. As Table 5 illustrates, most groups viewed coal as the primary energy source, followed by oil and then either natural gas, nuclear fission, or conservation. The anti-nuclear groups are the only dissenters, rating conservation as their top choice, solar heat as a major contributor and dismissing nuclear fission altogether. No other group ranks conservation higher than fourth, expects such a significant contribution from solar heat, or writes off nuclear energy altogether.

Table 5
RESOURCES THAT WILL MAKE MAJOR CONTRIBUTIONS TO OUR ENERGY NEEDS

Anti-Nuclear Groups	Industry	Financiers	NRC	Other Regulators	Congress	Outside Experts	
58%	96%	78%	94%	75%	91%	95%	Coal
50	57	50	63	84	67	79	Oil
42	44	41	34	48	52	67	Natural gas
0	52	24	28	52	25	33	Nuclear fission
42	1	0	3	12	10	2	Solar heat
100	16	29	19	50	52	38	Conservation

The researchers found it especially surprising that government agencies which frequently have been critical of nuclear energy were nevertheless looking to this resource for a large contribution. In fact, the other regulators' projection precisely matched that of the nuclear power industry.

An ongoing debate

The debate continues over the risks and rewards of nuclear electricity. In the heat of the debate, it's easy to get the impression that almost everyone has turned against nuclear electricity. This study shows that such an impression would be incorrect.

All energy sources have problems—environmental, safety, reliability, cost. Perhaps someday we'll find a perfect energy source, but until then we must provide for America's energy needs with sources we can count on today.

For a full report on the surveys discussed here, send for our free brochure, "Nuclear Electricity—Who Stands Where." Just fill out this coupon and mail it to:

U.S. Committee for Energy Awareness
P.O. Box 37012
Washington, D.C. 20013

NAME _____

ADDRESS _____

CITY _____ STATE _____ ZIP _____

225

Information about energy America can count on today

U.S. COMMITTEE FOR ENERGY AWARENESS

would assign funds for education and the improvement of facilities at research universities over the next four years. The fact remains that science as the search for new knowledge must continue to struggle for support in the U.S.

Casualties

The most sensitive indicator of the health status of a country or a locality, it is generally agreed, is the infant mortality rate: the number of infants dying within a year after birth per 1,000 live births. Recent UN estimates indicated an average infant mortality rate of 19 for the more developed countries and 100 for the less developed ones. In 1980 the U.S. rate was 12.5; the country ranked 19th, behind the United Kingdom and East Germany. Since then the overall rate for the U.S. has continued its long-term improvement, reaching a provisional 11.2 in 1982.

The rates for individual states and localities are something else again. Investigators of maternal and child health, state health officials and children's advocacy groups are deeply concerned to note that in many cases long-term rates of decline in infant mortality have flattened out and have even been reversed. They blame the recession, and some of them blame specific reductions in the funding of medical care by the Federal Government. Officials of the Reagan Administration argue that health-care cuts have not been significant, that no direct link can be established between the cuts (which took effect only in 1981) and any health-care outcomes and that the reported increases are "random" local and short-term fluctuations in generally improving trends.

Critics insist that Federal cutbacks have exacerbated the devastating impact of the recession on health care for people in need. The fraction of the U.S. population living below the official poverty level rose from 11.4 percent in 1978 to 14 percent in 1981. Millions of unemployed workers have lost their health insurance. In Michigan alone Blue Cross and Blue Shield lost 556,633 participants in less than three years. People without insurance seek free or low-cost care in community hospitals, many of which are threatened with insolvency, or turn to public programs for support. States and cities are hard-pressed financially, however, and Federal Medicaid payments were reduced by $1 billion in 1981. According to a survey made by the Children's Defense Fund, every state has reduced its Medicaid program for mothers and children by cutting back on services or making it harder to meet eligibility requirements; 44 states have specifically reduced prenatal and delivery services for pregnant women.

A clear circumstantial chain of cause and effect can be established between lack of health-care services and increased infant mortality. The key link is the low-birth-weight rate: the number of infants delivered weighing less than 2,500 grams (five and a half pounds) per 1,000 live births. A major correlate of low birth weight is inadequate prenatal care for mothers, and since 1980 increasing numbers of women have had too few prenatal-care visits or none at all. In Oregon the fraction of women who had no prenatal care doubled between 1980 and 1982, according to the Oregon Center for Health Statistics. The U.S. average low-birth-weight rate is some 70 per 1,000. (Japan's rate is 53 and Sweden's is 41.) In Oregon the low-birth-weight rate for 1979 through 1982 was 136.5 for women who had no prenatal care, about three times as high as for women receiving adequate care. In Michigan in 1978 the rate was 203 for women who had five or fewer prenatal visits and 57 for women who had six visits or more, according to a recent report from the Michigan Department of Public Health.

Low birth weight in turn increases the likelihood of lifelong handicapping conditions and appears to be associated with about two-thirds of infant deaths. Oregon data associate early death directly with lack of prenatal care: the 1979–81 neonatal mortality rate (death before 28 days) was more than nine times as high for infants whose mother had no prenatal care as it was for infants whose mother had adequate care.

Last winter an advocacy group, the Food Research and Action Center, reported (on the basis of a telephone survey) that infant mortality rates were rising in a number of states and urban areas, in most cases reversing the long-term improvement. Clearly not all the increases are statistically significant, as the center noted at the time. Administration spokesmen emphasized the continuing decrease in the overall U.S. rate in rebuttal. For large cities and states showing large increases, however, the data may indeed be significant. In a study to be published by the United Nations Children's Fund, C. Arden Miller of the University of North Carolina at Chapel Hill concludes that "to ignore [such reports] altogether suggests too much willingness to conceal within aggregate data some important clues about adverse effects on especially hard-hit people and locales."

Consider the case of Michigan, which has suffered three and a half years of double-digit unemployment, where 15 percent of the population receives some form of public assistance and where public-assistance grants have been cut repeatedly since 1979. Between 1950 and 1980 Michigan's infant mortality rate was reduced by 50 percent. It declined in every year from 1970 through 1980. In 1981 it rose by 3 percent, to 13.2, and last year it declined somewhat to a provisional 12.4. Regression lines calculated on the basis of total U.S. and Michigan rates for 1976 through 1980 highlight the difference between the aggregate data and those for an economically depressed state. The U.S. line predicts rates of 11.82 for 1981 and 11.18 for 1982, almost exactly the rates actually observed. The Michigan line predicts rates of 12.22 for 1981 and 11.68 for 1982 compared respectively with the observed 13.2 and 12.4. "In effect we lost a couple of years of progress," writes Jeffrey R. Taylor of Michigan's Division of Maternal and Infant Health; "229 more babies died than would have died if the downtrend had continued."

Taylor makes a point that is repeatedly emphasized by students of the subject: "Infant mortality has always been a problem of certain subpopulations in our society, not of the dominant white, well-insured, working majority. For an entire population's rate to go up, as in a large state such as Michigan (140,000 births per year), requires that certain subpopulations must be experiencing dramatic increases in suffering." Detroit's infant mortality is at about the total-U.S. level of the late 1960's. The city's rate was 21 in 1980, 21.9 in 1981 and provisionally almost 25 last year.

Health status in the U.S. is closely linked to socioeconomic status, Miller points out in his UNICEF report, perhaps because of "the nation's relatively weak commitment to assuring participation of all people in essential and appropriate health services." During the 1974–75 recession, however, there was a striking increase in health benefits, which "mitigated the adversities of unemployment and impoverishment." Now data are available "to lead reasonable policymakers to the inescapable conclusion that the health of children, pregnant women and poor families is suffering and in great jeopardy." The adverse effects must be attributed to "a combination of circumstances that include serious recession, increased poverty rates for households with children and diminished health benefits and social support services." In a time of local or widespread economic reversals, he points out, health services need to be expanded rather than contracted.

The Children's Defense Fund points out that building 239 MX missiles rather than the projected 240 would save $110 million—enough to fund Medicaid benefits for every pregnant woman living below the poverty level.

Microelectronic Projection

The advance of semiconductor technology over the past two decades has been reduced to strings of initials indicative of the growing density of microelectronic devices integrated on a

© 1982 Toyota Motor Sales, U.S.A., Inc.

Grand Touring. What does it mean?

In the new 1983 Toyota Cressida, Grand Touring means Grand Performance. With an amazing new 2.8 liter Twin Cam engine that does to smoothness and quietness of ride, what Michelangelo did to a ceiling. Add that to Cressida's new independent rear suspension, with coil springs and stabilizer bar, and its new electronically controlled 4-speed automatic overdrive transmission and you start to see how grand a driving experience can be.

Cressida's Grand Touring also means Grand Luxury. From the look of luxury outside — dashing, elegant, refined — to the feel of luxury inside — power windows and door locks. Automatic temperature control air conditioning. Variable assist power steering. Cruise control. AM/FM/MPX stereo receiver with four speakers. And a driver's seat that adjusts to your body in four distinct ways.

What else does Grand Touring mean?

In the new Cressida, it represents a feeling you get, while touring town or country. A feeling based on uncompromised performance. And spirited drive. It's the image of the car you've chosen. And of yourself.

Dashing.

BUCKLE UP—IT'S A GOOD FEELING!

OH WHAT A FEELING!

TOYOTA

CRESSIDA MOVES PROUDLY INTO THE GRAND TOURING CLASS.



DASH!



The IBM 4341 super-mini at universities

The BMOC, from engineering freshmen to Ph.D. researchers.

That's "Big Machine On Campus" at two big universities.

The University of Maryland

Engineering freshmen start using the IBM 4341 after only a few lectures. "Engineers will utilize the computer all their professional lives. The 4341 is so easy to use our students turn to it instinctively to solve problems," says Professor William Wockenfuss, who developed the introductory course.

Instructors are enthusiastic. "We know students use it. We get a rundown on how often they log in," says John Horan. "Log-in commands help them interact with the 4341 with no technical jargon," adds Aliakbar Elahi.



Dr. Glenn Ricart, Director of the Computer Science Center, believes, "With the 4341's 'Help' command, students pick themselves up if they stumble. That's critical in how they'll relate to computers in the future."

"The university's 4341s serve as a major resource for interactive computing," says James S. Harvison, Acting Assistant Vice President. "We have five 4341s installed at different campuses serving our university community. They are a definite asset to our instruction and research computing network."

The Johns Hopkins University

Here, the 4341 is used by The Center for Metropolitan Planning and Research. Teams of Ph.D.'s, under Dr. Steven Gottfredson, perform complex analyses of huge data bases. One study takes ten-year census figures and utilizes small samples to test the effects of federal programs on inner-city housing. Another breaks down census data to make it useful for planning by fire, sanitation and education departments.

Professor Gottfredson explains that "machine capacity and speed are vital because of the size of our data sets. We also handle many interactive algorithms, and the 4341 gives us fast access to information."

The 4341 is used in other areas at Johns Hopkins, including studies of cancer frequency and astrophysics applications in preparation for the planned launch of the space telescope. Vice Provost of the university, Richard Zdanis, says, "We decided on a 4341 because it's easy for inexperienced users and can support a variety of peripherals."

IBM systems, using microcode-assisted VM/CMS, help engineers and scientists in industry, research and universities think about the problem, not worry about the computer. We'd like to help you, too.

For an informative brochure, please write IBM Engineering and Scientific Marketing, 1133 Westchester Avenue, White Plains, N.Y. 10604.

IBM®

INVEST YOURSELF



A windmill to pump water for "salt farming" in India. More efficient woodburning stoves for the Sahel. Photovoltaic irrigation pumps for the Somali refugee camps

All these are solutions to technical problems in developing countries. Devising such solutions is no simple task. To apply the most advanced results of modern science to the problems of developing areas in a form that can be adopted by the people requires the skills of the best scientists, engineers, farmers, businessmen—people whose jobs may involve creating solid state systems or farming 1000 acres, but who can also design a solar still appropriate to Mauritania or an acacia-fueled methane digester for Nicaragua.

Such are the professionals who volunteer their spare time to Volunteers in Technical Assistance (VITA), a 20 year old private, non-profit organization dedicated to helping solve development problems for people world-wide.

Four thousand VITA Volunteers from 82 countries donate their expertise and time to respond to the over 2500 inquiries received annually. Volunteers also review technical documents, assist in writing VITA's publications and bulletins, serve on technical panels, and undertake short-term consultancies.

Past volunteer responses have resulted in new designs for solar hot water heaters and grain dryers, low-cost housing, the windmill shown above and many others. Join us in the challenge of developing even more innovative technologies for the future.

VITA Putting Resources to Work for People

1815 North Lynn Street, Arlington, Virginia 22209-2079, USA

single silicon chip. Beginning with SSI (small-scale integration) in the 1960's, the industry progressed to MSI (medium-scale integration) and LSI (large-scale integration) in the 1970's and is now entering the era of VLSI (very-large-scale integration). As the number of components on a chip has increased more or less exponentially with time (reaching into the hundreds of thousands in the case of some recently manufactured VSLI chips) the minimum dimensions of an individual circuit element (a line or a space) have shrunk dramatically: to a resolution approaching one micron in the most advanced designs. Given the inherent limitations of the photolithographic process employed to fabricate such devices, how far can this trend be expected to continue?

According to a leading investigator in the field, the prospects for reaching and breaking the "one-micron barrier" are good. In a review article on micron and submicron circuit engineering published in *Proceedings of the IEEE* (a publication of the Institute of Electrical and Electronics Engineers) Arnold Reisman of the Microelectronics Center of North Carolina forecasts that "by the end of the 1980's it is not unreasonable to expect that minimum lithographic dimensions will have decreased to the 1 μm [one micron] level" and that "by the end of the 20th century 0.25 μm minimum dimensions are anticipated." However, "based on what we know today, if this shrinkage occurs without major innovation in device, circuit and interconnection design, 0.25 μm may represent the end limit of usable linewidth shrinkage" in the fabrication of the transistors that now make up the bulk of the components in most integrated circuits.

For the immediate future the continuing miniaturization of microcircuit elements can apparently be accomplished by extending the "step and repeat" photolithographic process in widespread service today. In this process a "unit cell" pattern is projected in ultraviolet radiation through a mask onto part of a silicon wafer, and the wafer is repetitively stepped across the projector's field of view to expose the entire surface. As in earlier photolithographic methods, the radiation exposes a photosensitive emulsion, called a resist, which is then developed with a solvent; the unprotected areas of the silicon substrate are treated chemically to define a single circuit layer. Reisman reports that at present "there is considerable uncertainty about the minimum nominal dimension . . . achievable using optical exposure techniques for volume production." Nevertheless, "a somewhat soft consensus is that the optical resolution number is around 1–1.25 μm."

To achieve smaller dimensions, he predicts, shorter-wavelength radiation will be required. One possibility that has been under study for more than a decade is "direct writing" on the photoresist with a computer-controlled electron beam. Another is a step-and-repeat process utilizing X rays to expose the resist through precision masks generated by an electron-beam system. At present direct electron-beam writing is considered economically unfeasible for mass production because of its comparative slowness and high cost, but electron-beam systems are coming into widespread use for making photolithographic masks. In the long run, Reisman points out, the main problem with the shorter-wavelength approaches "may not be the fabrication of small structures but rather their reliability."

By the time one-micron minimum dimensions are achieved routinely in manufacturing it is estimated that a chip approximately six millimeters on a side will hold about a million components, corresponding to half a million bits of dynamic random-access memory, or DRAM. (Component counts are double DRAM counts, since every bit of information stored in such a memory requires both a storage capacitor and a transfer device.) At that stage, Reisman reckons, the cost of manufacturing such VLSI memory components will have fallen to "about .0002 cent per bit in 1983 money," or roughly $1 per chip.

Bedrock Pharmacopoeia

The pharmacopoeias of industrialized nations list on the order of 1,500 drugs. All have their uses, but only a few can be said to be basic to the health of a human population. Which are they? The question comes up because in many developing countries the system of health care is rudimentary, the supply of drugs depends on foreign suppliers and the ability to store and transport drugs in such a way that they do not deteriorate is limited. The World Health Organization Expert Committee on the Use of Essential Drugs looked into the problem and has published a list of 249 substances that would be appropriate in such conditions and a list of 22 substances, taken from the main list, considered essential for primary health care.

The committee based its choices on several criteria. One of them is that the drugs should meet the health-care needs of the majority of a country's population and therefore should "be available at all times in adequate amounts and in the appropriate dosage forms." Another is that a given country's choice of drugs will depend on many factors, including the pattern of diseases prevalent there, the facilities for treatment and the training and experience of health workers. Moreover, "only those drugs should be selected for which sound and adequate data on efficacy and safety are available from adequate clinical studies and



We're light on your feet.

Our casting material is up to one-third lighter than plaster. It's three times stronger and we hope you never need it.

But on most any injury requiring a cast, you'd find Scotchcast® Casting Tape more comfortable than plaster and easier to get around in. It's thinner, so it fits under clothing. It also breathes. That reduces itching, odor and helps you stay cooler in warm weather.

All the comforts of Scotchcast tape were inspired by a 3M employee who asked for something "less barbaric" than his plaster cast.

Hearing the needs of the medical profession and the patients it serves has helped 3M pioneer over 700 health care products. We now make everything from stethoscopes and pharmaceuticals to x-ray film that cuts radiation danger in half.

And it all began by listening.

3M hears you...

For your free 3M Products and Capabilities Brochure, write: Department 010709/3M, P.O. Box 22002, Robbinsdale, MN 55422.

Name_____

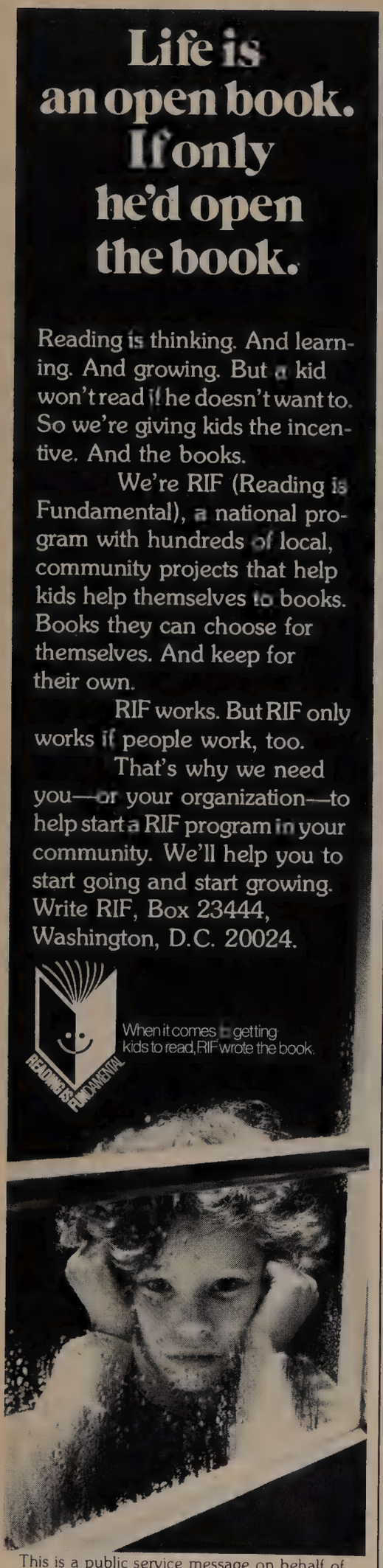
Address_____

City_____ State_____ Zip_____

Or call toll-free: **1-800-33M-HEAR** or 1-800-336-4327.

©1983, 3M

3M



This is a public service message on behalf of Reading Is Fundamental and this magazine.

for which evidence of performance in general use in a variety of medical settings has been obtained."

Applying these criteria, the committee made up its main list in 27 categories. Among them are analgesics; antidotes; anti-infective drugs; blood products and drugs affecting the blood; drugs for cardiovascular, dermatologic and gastrointestinal maladies; hormones; muscle relaxants; ophthalmological preparations; psychotherapeutic drugs; drugs acting on the respiratory tract, and vitamins and minerals.

The 22 substances on the committee's "model list of drugs for primary health care" are acetylsalicylic acid (aspirin), activated charcoal, an antacid, an antihemorrhoidal drug, atropine (an antispasmodic), benzoic acid plus salicylic acid, benzyl benzoate, calamine lotion, chlorhexidine solution, chloroquine, chlorphenamine, ephedrine (for asthma), ergometrine (for bleeding after childbirth), iodine, ipecacuanha, iron and folic acid (a nutritional supplement during pregnancy), lindane, mebendazole, oral rehydration salts, paracetamol, piperazine and tetracycline eye ointment. The committee said the drugs on this list "can be used effectively and safely by responsible individuals with little formal medical knowledge."

Conjugate Bliss

Light is notoriously subject to aberrations induced by the medium through which it passes; indeed, the science of optics came into existence largely to correct them. Now a novel kind of mirror optics is emerging in which aberrations are corrected by the light itself. The mirrors are called phase-conjugate mirrors. The first such mirror was developed in the early 1970's by B. Ya. Zel'dovich and his colleagues at the Lebedev Institute in Moscow. It was simply a tube of methane under pressure. Since then workers at the Hughes Research Laboratories, the California Institute of Technology, Bell Laboratories and elsewhere have explored more complex systems of beams and materials, notably crystals.

In such a system a beam of laser light passing through a distorting medium (for example the atmosphere, which is heated by the beam and acts as a dispersing lens) is made to strike the phase-conjugate mirror. The mirror, which is solid, liquid or gaseous, is a material whose atomic structure is maintained by binding forces not markedly greater than the electric field strength of the beam. Hence the beam affects the structure. In particular the beam affects the material's index of refraction, thereby changing its own propagation. In effect a mirror develops inside the material, and so the beam is reflected.

The reflection, however, is curious. For one thing, the reflected beam returns along the precise trajectory of the incident beam. With an ordinary mirror this happens only if the beam arrives perpendicular to the mirror. Then too the reflected beam is the phase conjugate of the incident beam. That is, the evolution of the spatial information in the reflected beam (for example the spreading of the beam caused by the atmosphere) reverses the evolution of the information in the incident beam. Thus the return of the beam through the medium that distorted it restores the beam to its initial, undistorted state.

The reversal of distortion promises several applications. In a scheme for controlled nuclear fusion, fuel pellets are to be heated with lasers; phase-conjugate mirrors might keep the pellets well targeted. In a scheme to increase the power and speed of computers, circuits that manipulate electrons are to be superseded by circuits that manipulate light; phase-conjugate mirrors might make the circuits self-correcting. In photolithography, integrated circuits are manufactured by etching a photographic emulsion after parts of it have been exposed to light; phase-conjugate mirrors might make the process more precise.

A system of phase-conjugate optics being developed by Amnon Yariv and his colleagues at Cal Tech is intended to serve communication channels by eliminating the requirement that laser beams make two-way trips, one to reach a phase-conjugate mirror and one to return. In their arrangement a "probe beam" and an "information beam" travel in opposite directions. The probe beam leaves the receiving station of a communications link and travels toward the broadcasting station, becoming distorted on the way. On its arrival at the broadcasting station it affects a phase-conjugate device, so that the device can give the information beam a distortion complementary to the one the probe beam has undergone. The information beam can then take flight in a predistorted form. The distorting medium through which it will pass will now serve to correct its information, so that it arrives distortion-free.

Waiting for Halley

The apparition of Halley's comet in 1910 was so spectacular that, photographed by the largest telescopes of the time, it became a kind of astronomical allegory. Soon after the comet's closest approach to the sun it had a bright coma (the gaseous and dusty cloud that envelops the nucleus of a comet) and a luminous double tail extending across a fourth of the sky. What will the comet look like on its return in 1985 and 1986? Writing in *The Journal of the Royal Astronomical Society of Canada,* Ian Halliday

ATTEND TELECOM 83 GENEVA

Switzerland—October 26-November 1, 1983
Fourth World Telecommunication Exhibition

TELECOM 83, the fourth in a series of quadrennial World Telecommunication Exhibitions, is organized under the auspices of the International Telecommunication Union (ITU) and its 158 member countries.

Highlighting the general theme "Telecommunications for Everyone," TELECOM 83 offers a unique opportunity for individuals or groups to join in the exchange of ideas, information, and technology in the fields of telecommunications and electronics. Additionally, it will provide an excellent setting for international contacts.

An integral part of TELECOM 83 will be the Fourth World Telecommunication Forum, organized by the ITU and 50 national and international engineering societies from all five continents. Telecom and the World Telecommunication Forums are recognized as the universal and most authoritative meeting of telecommunication engineers and economists.

FORUM 83: The World Telecommunications' Summit

FORUM 83—Parts 1, 2 and 3—will assemble a "brains trust" of several thousand top executives who will present and discuss the planning, financing, management and implementation of the world telecommunication network, and the convergence of computing and communications technologies.

Over 200 presentations will enable participants to hear, meet and question industry and government leaders on current and future developments of the world telecommunication network. Plenary sessions and selected parallel sessions will have simultaneous interpretation.

Nations Working Hand in Hand

The three-day Part One of FORUM 83 will offer presentations by prominent figures in the economic sector as well as the telecommunication field. Speakers will include government leaders, senior corporate managers, chief scientists from industry, and representatives from international and financial organizations. They will discuss the technological requirements of industrialized and developing nations. They will also confront the future need for financing national, regional and international telecommunication development plans.

Towards a World Network

FORUM 83 Part Two has adopted the theme: "Integration of the World Telecommunication Network." This three-day event will be a far-reaching technical and scientific symposium with wide-ranging appeal. It represents the combined efforts of the ITU's 158 member countries to introduce to those attending, the advances made since FORUM 79 to solidify the world telecommunication network. Some 60 presentations will be made covering the latest innovations and technological trends.

Participation for each of these parts of FORUM 83 is limited to the first 1500 registrants.

In keeping with the ever-increasing transnational exchange of data, information and broadcasting, FORUM 83 Part Three, a Legal Symposium, will examine the legal aspects of international telecommunication. This concluding session of TELECOM 83 will scrutinize the international regulations relating to transnational information transport.

TELECOM 83 Exhibition

The Telecom 83 Telecommunication Exhibition encompasses all of the exhibition, conference and outdoor floorspace of Geneva's New Exhibition and Conference Centre. This quadrennial Telecommunication Exhibition will be a kaleidoscope showcasing the world's telecommunications and electronics industries, their capabilities and their networks.

The BOOK FAIR at TELECOM 83 will provide an opportunity for delegates to reference current works on telecommunications, electronics and allied fields.

Registration Fees (Swiss Francs)

	Individual	Group*	Officials**
Part 1	600	550	500
Part 2	300	275	250
Part 3	250	250	250
Parts 1 and 2	750	750	750
Parts 1 and 3	750	750	750
Parts 2 and 3	500	500	500
Parts 1, 2 and 3	850	850	850

*two or more from the same company or organization and members of institutions co-sponsoring the Technical Symposium.
**official representatives from administrations of countries/members of the ITU.

TELECOM 83 **Registration Form**

TO: **FORUM 83**
International Telecommunication Union
Place des Nations
CH-1211 Geneve 20, Switzerland

Please register me for the following Forum '83 session(s).

Part 1 ☐ at rate for individuals ☐ groups ☐ officials ☐
Part 2 ☐ at rate for individuals ☐ groups ☐ officials ☐
Part 3 ☐ at rate for individuals ☐ groups ☐ officials ☐

Name _____

Organization/Job Title _____

Address _____

_____

I enclose a cheque for SwFr of _____ in full payment of my registration.

I have initiated a bank transfer of SwFr _____ to Credit Suisse Geneva a/c Wagonslits Tourisme Cooks TW 673.050.71. Specify TELECOM 83 Forum.

Date _____ Signature _____

The Oceanic Crust

It is created and destroyed in a flow outward from midocean ridges to subduction zones, where it plunges back into the mantle. Currently it is being opened to view by submersibles and novel instrumentation

by Jean Francheteau

From the point of view of the earth scientist our planet probably should be called Ocean rather than Earth, not only because 70 percent of it is covered by water but also because 60 percent of its solid surface is covered by the thin crust that is manufactured in a unique geologic mill at midocean. The first piece of the oceanic crust to be identified was recovered by the British cable-laying steamship *Faraday* in 1874. The *Faraday* was sailing in the North Atlantic on a mission to mend a broken transatlantic telegraph cable. The cable had broken in 2,242 fathoms of water where it passed over a large rise in the ocean floor later named the Faraday Plateau. The *Faraday* was equipped with a large grapnel for lifting cables from the ocean floor.

According to Marshall Hall, an English geologist, "whilst engaged in grappling for the broken telegraph cable the ship caught the strong claws of the grapnel in a rock, which resisted with a strain of about 27.5 tons, to which any but a rope of marvelously perfect manufacture would have yielded. As it was, the rock gave way and a lump of black basalt came up weighing 21 lb. This mass shewed signs of having been torn off." The basaltic rock was brought back to England, and in 1876 Hall and J. Clifton Ward examined and described it.

In the century following the finding of this first piece of it the oceanic crust has come to have a central position in earth science. It is now known that the Faraday Plateau is one segment of a 59,000-kilometer system of ridges that girdle the earth under the sea. The midocean ridge has great significance in the theory of plate tectonics, which transformed the earth sciences in the late 1960's and early 1970's. The ridge marks the boundary between two rigid plates that are supported by the underlying mantle of the earth. At the ridge the plates separate slowly and the underlying rock rises to fill the gap, melting as it does so. Thus a few square kilometers of new oceanic crust is formed each year at the crest of the midocean ridges. The crust that is formed in this way is profoundly different from the crust of the continents. It is an order of magnitude younger on the geologic time scale than the continental crust and has a quite different composition from the continental masses.

As a result of many decades of observation that were unified in the plate-tectonic hypothesis it is now known where and roughly how oceanic crust is formed. The detailed structure of the crust, however, is much less well known. Therefore in the late 1970's and early 1980's the work on the oceanic crust has turned from global theory-making to investigating details of structure and composition. The oceanic crust now appears to be much more diversified both in its topography and in its layered structure than had previously been thought. The refinements in understanding the oceanic crust that have been achieved are due in large part to novel techniques of observation of the sea floor, which remains one of the least accessible parts of the planet's surface.

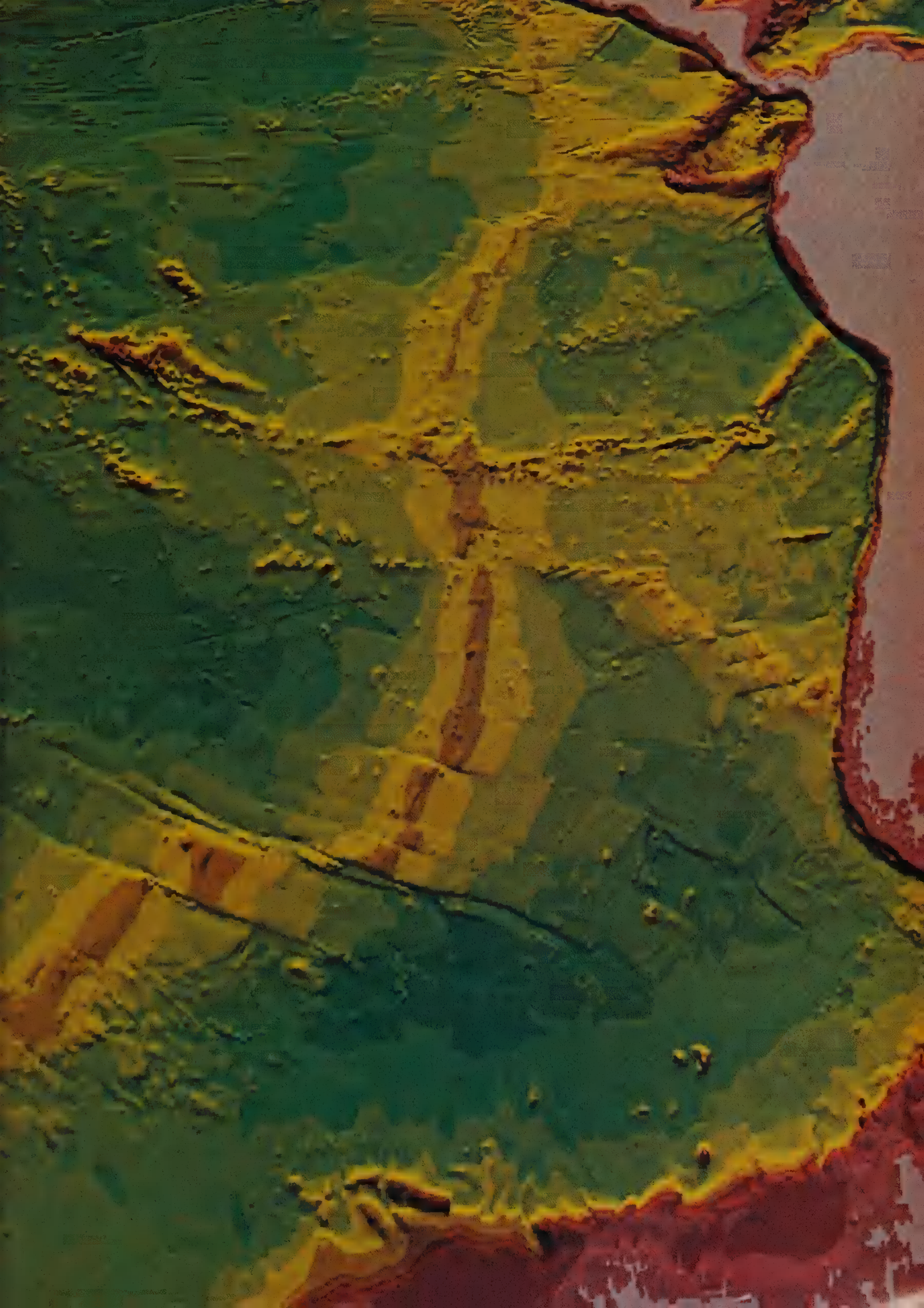
Satellite measurements of the gravity field over the oceans are yielding an improved picture of the general topography of the ocean bottom. The detailed topography is being mapped by new sonic methods. The layers of the crust are being probed by deep drilling, by new seismic methods and by measurements of electrical conductivity. The large body of new information that is being made available by such techniques is rapidly changing the accepted notion of the oceanic crust. The work is by no means complete, but within a few years we should have a substantially better picture of the thin crust that covers the larger part of the earth's solid surface.

In plate-tectonic theory the crust and the upper mantle of the earth are divided into the lithosphere, or strong layer, and the asthenosphere, or weak layer. The lithosphere includes the crust and part of the upper mantle. In the ocean the crust is on the average from five to seven kilometers thick; away from the ridge crest the lithosphere is about 100 kilometers thick. The lithosphere is broken up into a set of fairly rigid plates that are much like rafts floating on the less rigid material of the asthenosphere.

The plates move at a rate of a few centimeters per year with respect to each other, and the boundary between two plates can be described according to the relative plate motion. At divergent boundaries the plates separate. At convergent boundaries the plates move toward each other, and one plate generally plunges under the other and into the asthenosphere in the process called subduction. At transform boundaries the plates slide past each other. The spreading center of the midocean ridge, where molten rock from the mantle is injected into the crust, is a divergent plate boundary.

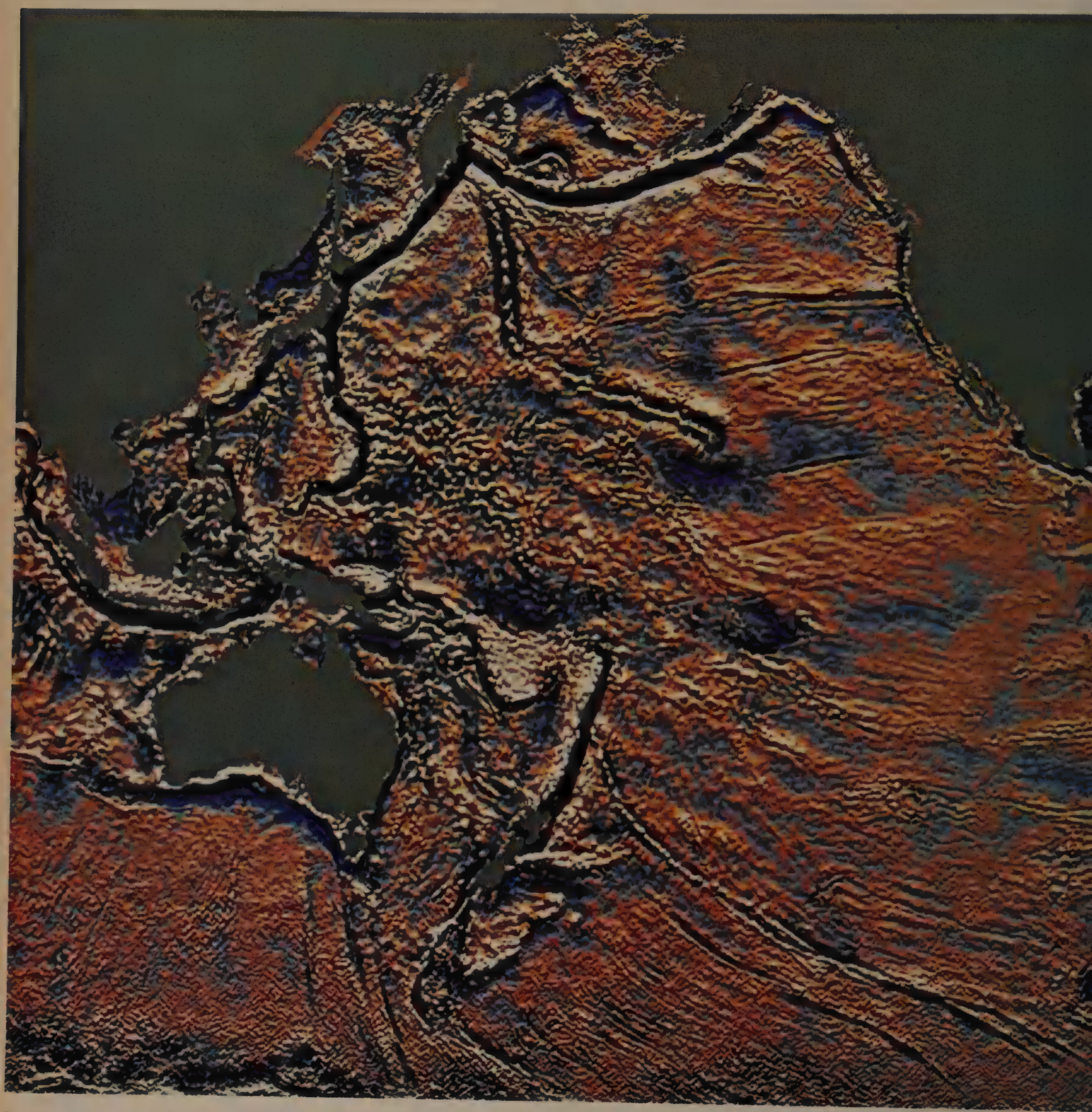
The mantle below the spreading center is composed mainly of peridotite, a type of rock consisting mainly of the mineral olivine, which in turn consists

EAST PACIFIC RISE appears as a raised area running roughly north and south in the topographic map on the opposite page showing the Pacific Ocean off the coast of South America. Brown designates shallow regions, yellow intermediate regions and green deep regions. The East Pacific Rise is part of a ridge system that circles the earth under the sea. Each ridge marks the place where two lithospheric plates diverge: the East Pacific Rise is the boundary between the Pacific plate and the Nazca plate. As the plates separate, molten material wells up to fill the gap and creates new oceanic crust. The axis of the rise is crossed by many of the large faults called transform faults. The map is based on depth soundings made by research vessels; the data were compiled by the U.S. Naval Oceanographic Office. The map was plotted and colored by a computer at the Lamont-Doherty Geological Observatory of Columbia University.



mainly of magnesium, iron, oxygen and silicon but is poor in silicon compared with the rocks of the crust. The separation of the plates at the ridge reduces the downward pressure on the mantle rock below. Hence a part of the mantle begins to move upward; the zone of upwelling extends from a depth of 50 to 70 kilometers to the base of the crust. The decompression of the mantle material is adiabatic, meaning that it takes place without loss of heat, and under such conditions the peridotite begins to melt as it rises.

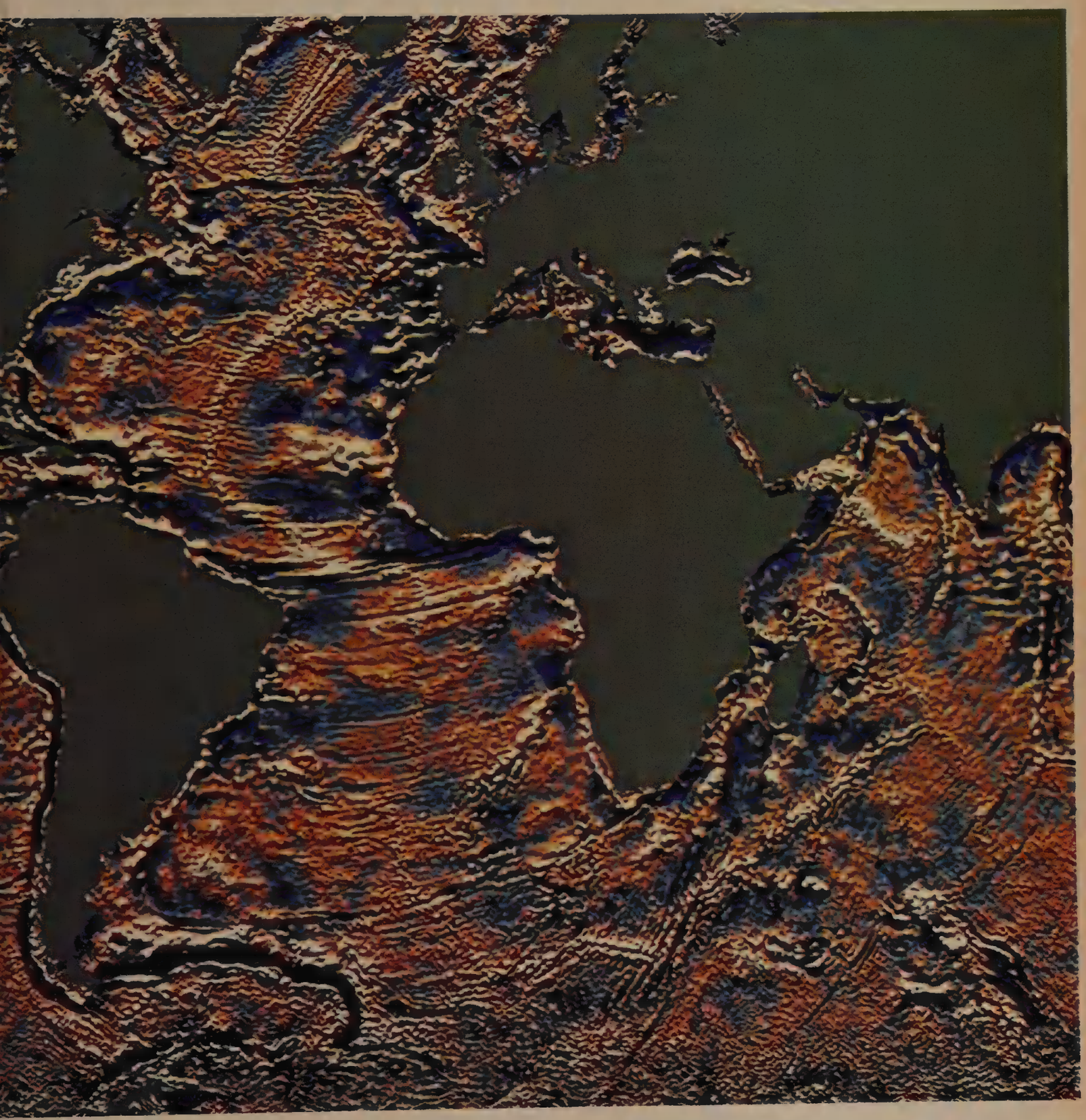
Not all the peridotite melts on the way to the surface. In general the basaltic liquid injected at the spreading center is formed by the melting of 10 to 20 percent of the upwelling mantle rock. The melt collects in a magma chamber at the base of the crust, where it separates through crystal fractionation into fractions of different composition. The fractions lie above a residual solid that has the composition of peridotite. Within the chamber slow cooling and crystal fractionation result in the formation of



GEOTECTONIC IMAGERY based on satellite gravity measurements reveals the tectonic features of the ocean floor as if the seas had been drained. White areas correspond to the high gravity fields associated with shallow regions of the ocean floor, blue areas to the low gravity fields associated with deep regions and red areas to intermediate gravity fields. The raised red-and-white strip between Europe and America marks the Mid-Atlantic Ridge. The dark blue lines in the western Pacific show the effect of deep oceanic trenches; the trenches are surrounded by shallow areas. The irregular diagonal line in the Pacific is the Emperor-and-Hawaii seamount chain. A

gabbro, a type of rock incorporating in addition to olivine the mineral plagioclase, which consists mainly of silicon, oxygen, sodium and calcium. Gabbros and other rocks that are formed by accumulation make up the lower layer of the crust. The basaltic liquid in the upper part of the chamber reaches the surface through a system of vertical passages. Once at the surface, the liquid flows down the slopes of the ridge and hardens into sheets or the rounded forms known as pillow lavas. Which of the forms results depends on the slope of the ridge and the rate at which the lava is extruded. Moreover, as the plates move apart the ascending magma hardens into a series of dikes: massive vertical sheets.

The injection of magma, or molten silicate liquid, plugs the gap left by the moving apart of the plates. The plates continue to diverge, however, and the plug is ultimately rifted open. A new cycle of asthenospheric upwelling, peridotite melting, separation in the magma chamber and extrusion begins. Meanwhile the crust formed in the previous



seamount is an undersea volcano; Hawaii is at the southern end of the chain. The map is based on measurements of the height of the sea surface made by *Seasat*, a satellite launched by the National Aeronautics and Space Administration in 1978. The computer methods for making geotectonic images were developed by William F. Haxby of Lamont-Doherty. Differences in the mass of the rock on the sea floor result in variations in gravity: seamounts have a high gravity field and trenches have a low field. Hence water tends to "pile up" over seamounts and to do the opposite over trenches. Therefore the ocean bottom can be inferred from the height of the sea surface.

round of upwelling is moving outward from the spreading center. As the crust moves outward it is modified. Tension exerted by the continuing plate motion can result in a series of fissures and faults running parallel to the strike, or longitudinal axis, of the ridge crest.

As the crust cools many small cracks appear in its upper layers. Seawater penetrates the cracks and the fissures caused by the tension in the crust. The water flows downward into the crust, is heated and rises to the surface of the crust again. Such hydrothermal circulation efficiently leaches water-soluble compounds out of the rock. Metallic elements, which readily form ionic complexes, are leached with particular efficiency. The leached elements are carried upward and the hot seawater with its burden of metals is expelled into the ocean by vents near the ridge crest. The discovery of the vents and the exotic biological communities that cluster around them was one of the most exciting findings of oceanic science in the late 1970's.

The hot circulating water and shallow heat sources lead to metamorphic changes in the lower crust and rapid chemical changes in the upper crust. In addition, as the crust moves outward a layer of sediment is deposited on its upper surface. The sediments consist largely of the remains of tiny oceanic plants and animals. The type of sediment that is deposited and the depth of the sediment cover therefore depend largely on the biological activity of the ocean.

Thus at some distance from the ridge crest the oceanic crust has the following vertical structure from the ocean floor downward. At the top is a layer of sediment perhaps .5 kilometer thick. Under the sediment is a layer known as the oceanic basement made up of interspersed sheet flows and pillow lavas with the underlying complex of vertical dikes. The basement can be two kilometers thick and is heavily fractured and altered by hydrothermal circulation. The third layer, the oceanic layer, is made up of the gabbros that solidify out of the basaltic melt in the magma chamber. The gabbros can undergo considerable metamorphism as they move away from the ridge. The oceanic layer is perhaps five kilometers thick.

This model of the layered structure of the oceanic crust and how it is formed is based on marine geophysical data, on studies of rocks from the ocean floor, on observations of fragments of oceanic crust emplaced in mountain belts on dry land and on conjectures. The rest of this article will be devoted to examining some of the unsolved problems presented by that model. It will become clear that the simple picture of the crust that was thought to be accurate until quite recently is rapidly being made more complex by new findings.

Since the first deep-sea soundings were made in the second half of the 19th century it has been known that the ocean floor lies much deeper below the sea surface than the continents rise above it. Away from the continental margins, which are not made up of oceanic crust, the ocean is on the average 3.7 kilometers deep. The great depth of the ocean and the sediment cover on the oceanic crust make the crust difficult to observe with most geologic techniques. In the past decade, however, several technical advances have greatly increased the accumulated knowledge of the oceanic crust.

At the most general level the methods of observation can be put in two groups: those that reveal the topography of the surface of the crust and those that penetrate below the ocean floor to gain information about the composition and vertical structure of the crust. In the studies of topography much interest is focused on the midocean ridge that marks the spreading center. The ridge is a long linear upswelling with a gradual slope. The average depth at the crest of the ridge is about 2.5 kilometers, and the ocean bottom slopes away on both sides to a depth of from five to six kilometers.

As the oceanic crust is carried away from the ridge crest it cools and contracts. In this process the lithospheric plate can be thought of as "floating" on the asthenosphere. Consider a block of wood floating in a barrel of water. If the block is not perturbed, it will come to rest in the water at a level corresponding to a balance between the downward force of gravity and the buoyancy of the wood. The buoyancy and hence the point of equilibrium depend on the density of the wood. Similarly, in the absence of perturbing factors the lithospheric plate floats on the asthenosphere at a depth corresponding to what is called isostatic equilibrium: the level at which the weight of the lithosphere is balanced by the pressure from the mantle. The level to which the plate sinks depends on the density of the rock in the lithospheric column, a thin vertical section through the lithosphere.

As the oceanic crust cools and contracts its density increases and hence it sinks deeper into the asthenosphere. The depth to which the crust sinks has been shown to vary with the square root of its age. Crust that is two million years old lies at about three kilometers, crust that is 20 million years old lies at four kilometers and crust that is 50 million years old lies at five kilometers. Charting the topography of the ocean bottom can therefore provide an estimate of the age of the crust.

Topographic maps can yield other significant information on the movement of the plates. Because lithospheric plates are rigid bodies, when two plates separate, their motion can be described in relation to a point on the earth's surface that is referred to as the pole of rotation. (The pole of rotation should not be confused with the earth's geographic or magnetic poles; it is significant only in relation to plate motion. Furthermore, the pole describing the relative motion of a pair of plates can shift several times over the history of the plate's interaction.) Continuous segments of spreading axis along the ridge crest define great circles that pass through the pole of rotation much as a meridian, or line of longitude, passes through the geographic pole.

As two plates move around their pole



of rotation transform faults caused by stresses in the plates form across the strike of the ridge. Other transform faults are caused by the irregularity of the initial rupture within a continent that created the ocean basin. When the land masses separate, the breaks parallel to the direction of plate motion develop into transform faults; the breaks perpendicular to the plate motion develop into spreading centers. At a transform fault the axis of the ridge is offset, and the parts of the plates slide by each other in opposite directions.

The offsets at the transform faults are roughly parallel to the direction of relative plate motion. The offsets have the effect of keeping the axis perpendicular to the direction of spreading. On the Mid-Atlantic Ridge transform faults can be as close together as 50 kilometers. Away from the ridge, outside the offset, the parts of the plate are not in motion with respect to one another. In this region structures called fracture zones mark the position of the transform faults; the fracture zones extend from the ridge crest like ribs from a spine. Between the transform faults are many smaller faults that are also caused by plate motion.

The precise position of the faults and fracture zones on the ridge holds an excellent record of the kinematics of the plates: the history of the relative plate motions. Mapping techniques that enable geologists to identify the position of these transverse features can therefore serve as the basis for reconstructing the history of the plates, by a method analogous to projecting a reel of motion-picture film in reverse.

One of the most promising tools for reconstructing the kinematics of the plates is a kind of mapping that has been named geotectonic imagery by its developer, William F. Haxby of the Lamont-Doherty Geological Observatory of Columbia University. Geotectonic images come from data gathered by *Seasat*, a satellite launched by the National Aeronautics and Space Administration in June, 1978. *Seasat* was equipped with a radar altimeter capable of measuring the height of the sea surface to within five to 10 centimeters. The instruments on board the satellite failed prematurely after three months of operation, but by then the device, orbiting at an alti-



MAJOR LITHOSPHERIC PLATES and their boundaries are mapped. The lithosphere consists of the crust and the rigid upper part of the mantle. In midocean, away from the spreading center, the lithosphere is about 100 kilometers thick. The upper five to seven kilometers is crust. The lithospheric plates are in motion with respect to one another, and the boundaries are defined according to this relative motion. At divergent boundaries the plates move apart; the spreading centers at the midocean ridges are divergent boundaries. At transform faults the plates slide past each other. At convergent boundaries the plates move together and one plate plunges under the other in the process of subduction. Thus oceanic crust is created at divergent boundaries and destroyed at convergent boundaries.



WWF/Nancy Nash

The panda stands for WWF and for thousands of other animals and plants facing extinction

THE WORLD WILDLIFE FUND (WWF) is dedicated to the conservation of all endangered forms of life. Sadly, the Giant Panda is one of the many species now in danger of extinction.

In a unique and historic example of international co-operation the People's Republic of China have invited WWF to work with them to save the world's most widely-admired animal.

The Chinese Government has been actively engaged in Panda Conservation for many years. Now a WWF team led by the distinguished ecologist Dr. George Schaller is at work in Sichuan Province together with top Chinese scientists under the leadership of Professor Hu Jinchu to carry out an in-depth study of the Panda and its needs for survival in the wild.

A major problem: the Giant Panda's diet demands huge quantities of bamboo and the evidence suggests that the bamboo in Wolong Natural Reserve may be about to flower and die – a serious threat to the survival of Pandas in that area.

Other factors – the Panda's low reproduction rate, internal parasites, dietary and territorial requirements – are also being studied.

Ultimately, to ensure that the Giant Panda has a future, we have to conserve the complex ecosystem in which it lives. This broad conservation philosophy is reflected in the hundreds of well-planned projects which are now being carried out by WWF in over 50 countries.

The Giant Panda is an endangered animal. It is also the symbol of WWF's worldwide conservation efforts to save life on earth.

But WWF needs money – your money.

Please send your contribution to:
WWF International,
1196 Gland, Switzerland
Or to the WWF National Organisation in your country.

I support the aims of the World Wildlife Fund and enclose the sum of

_____

Name_____

Address_____

_____

_____ SA

WWF© WORLD WILDLIFE FUND

WWF acknowledges the donation of this space by Scientific American.

Advertisement prepared as a public service by Ogilvy & Mather.



Dance has a new partner.



Business. 200 corporations know that dance is important to the people important to them. That's why they are investing in seven of America's greatest dance companies through The National Corporate Fund for Dance. Don't let your corporation sit this one out. Contact William S. Woodside, Chairman, American Can Company c/o The National Corporate Fund for Dance, Inc., 130 West 56th Street, New York, N.Y. 10019.

THE NATIONAL CORPORATE FUND FOR DANCE

tude of 800 kilometers, had surveyed the world's oceans between 72 degrees north and 72 degrees south.

The data from the three months of *Seasat* operation are currently being processed to yield dramatic images of the ocean floor [*see illustration on pages 116 and 117*]. The principle employed to convert the information on sea height into topographic maps is an intriguing one. The main cause of the spatial variation in the height of the sea surface is variation in the gravity field at sea level: the ocean tends to "pile up" where the gravity field is high and do the opposite where the gravity field is low.

The differences in gravitation are measured in relation to the ellipsoid. The ellipsoid is a mathematical figure corresponding to what the average sea surface would be if the mass of the earth were distributed in a radially symmetrical way.

The mass of the upper layers of the earth under the oceans is not, however, distributed in a radially symmetrical way. Hence the sea surface does not follow the ellipsoid. Instead it follows an irregular figure called the geoid. The difference between the geoid and the ellipsoid at any point on the ocean corresponds to the gravitational anomaly there. Where the geoid is higher than the ellipsoid there is a positive gravitational anomaly. Where the geoid is lower there is a negative anomaly.

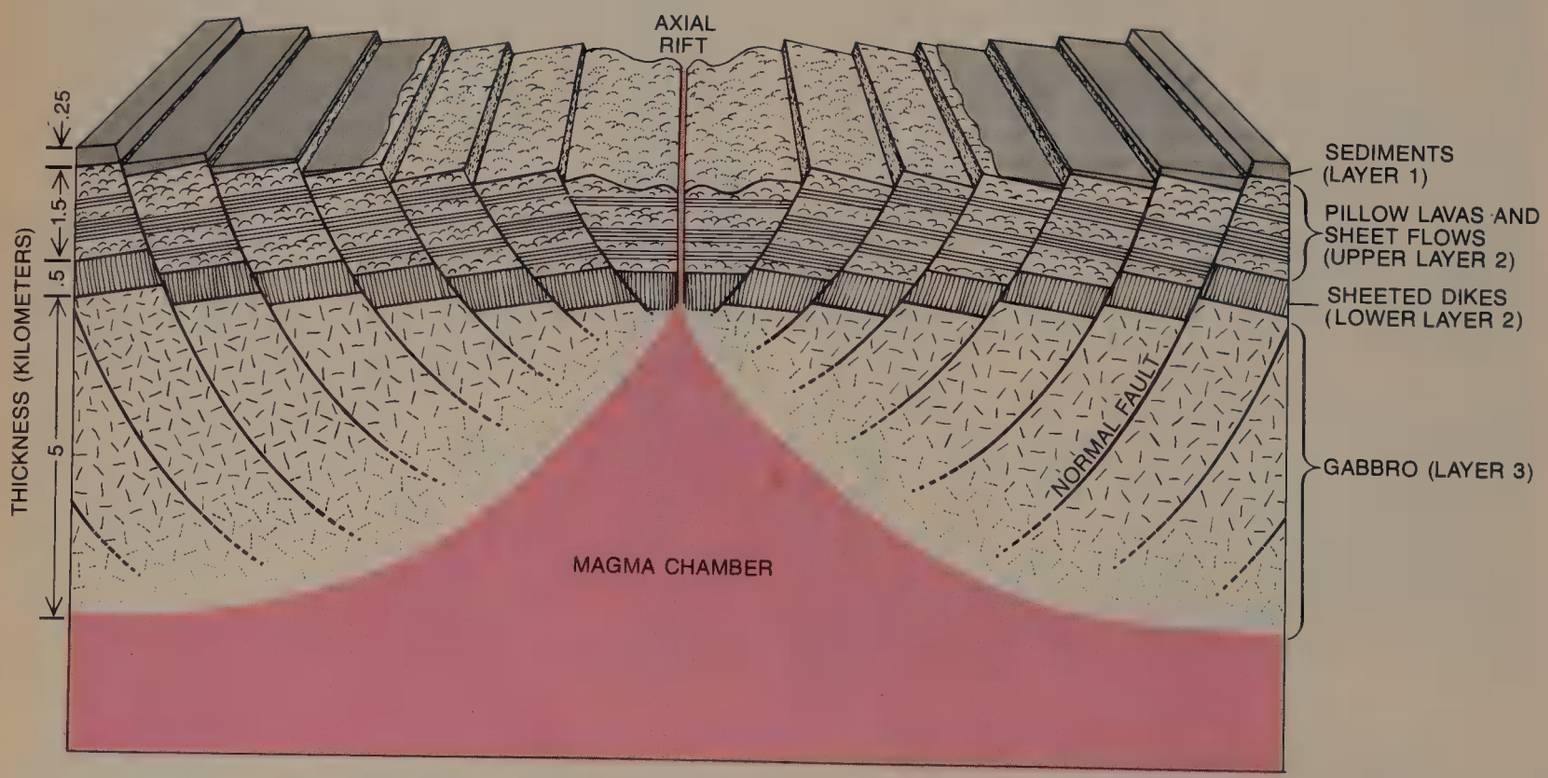
What could account for the differences in the gravity field? Since the gravitational force exerted by a body is proportional to its mass, the presence of large topographic features such as seamounts, or undersea mountains, rising from the ocean floor is associated with positive gravitational anomalies; depressions or valleys in the floor are associated with negative anomalies.

In the *Seasat* maps of the ocean floor the fracture zones, volcanic seamounts and deep trenches near subduction zones stand out as if the ocean had been drained away. Their spectacular vistas aside, geotectonic images are being put to significant geophysical uses. For example, by providing detailed maps of fracture zones the geotectonic images are helping in the reconstruction of the direction of plate motion in the past.

The geotectonic images do not correspond exactly to the topography of the sea floor. Although seamounts, fracture zones and trenches are rendered with great clarity, one of the main spreading centers in the Pacific, the East Pacific Rise, appears as an insignificant swell. The reason for the discrepancy can be understood by considering again the block of wood. Imagine the block is a topographic feature on the crust and *Seasat* is flying over it comparing its gravity field to the fields around it.

When the block floats, part of it extends above the surface; hence there is an additional mass at that point on the surface and a positive gravitational anomaly would be expected there. The block is less dense than the fluid in which it floats, however, and so below the surface there is somewhat less mass than there is at the nearby points where no blocks are floating. When the block is at flotational equilibrium, the mass added above the surface and the mass subtracted below the surface are equal. Therefore under conditions of isostatic equilibrium in the crust a much smaller gravitational anomaly is observed than would be expected on the basis of the topography alone. The topographic feature is thus almost invisible to *Seasat*.

The midocean ridges are approximately in isostatic equilibrium, and so they appear as relatively small upswellings, much smaller than they are in reality. Other features are out of equilibrium and appear quite clearly. Consider a seamount on a piece of old oceanic crust. The crust has cooled and contracted and is quite rigid; hence the seamount cannot sink down far enough to reach equilibrium at that point. Instead the crust is depressed only slightly under the seamount but the shallow depression extends a long distance around the mount. Thus the large area that includes the seamount is in equilibrium but the point where the mount sits is not. The subtracted mass is spread over a considerable area of the crust, whereas the added mass is concentrated at the



SPREADING CENTER is where magma is injected into the crust, as is shown in this cross section of the midocean ridge. The magma forms as the lithospheric plates diverge and mantle rock rises and melts with the decrease of pressure. It collects in a chamber below the spreading center. Within the magma chamber the rock gabbro crystallizes. At the top of the chamber the magma rises as the plates diverge and cools as vertical dikes. At the surface lava flows out and hardens in the form of sheets and "pillows." As the new crust moves away from the spreading center, a layer of sediments is deposited on it. The crust also cracks along normal faults, which run parallel to the ridge crest. Hence the mature crust has a layered structure from the top down: sediments, sheet and pillow lavas, dikes and gabbros.

seamount. The result is a positive gravitational anomaly at the seamount and a smaller negative anomaly around it.

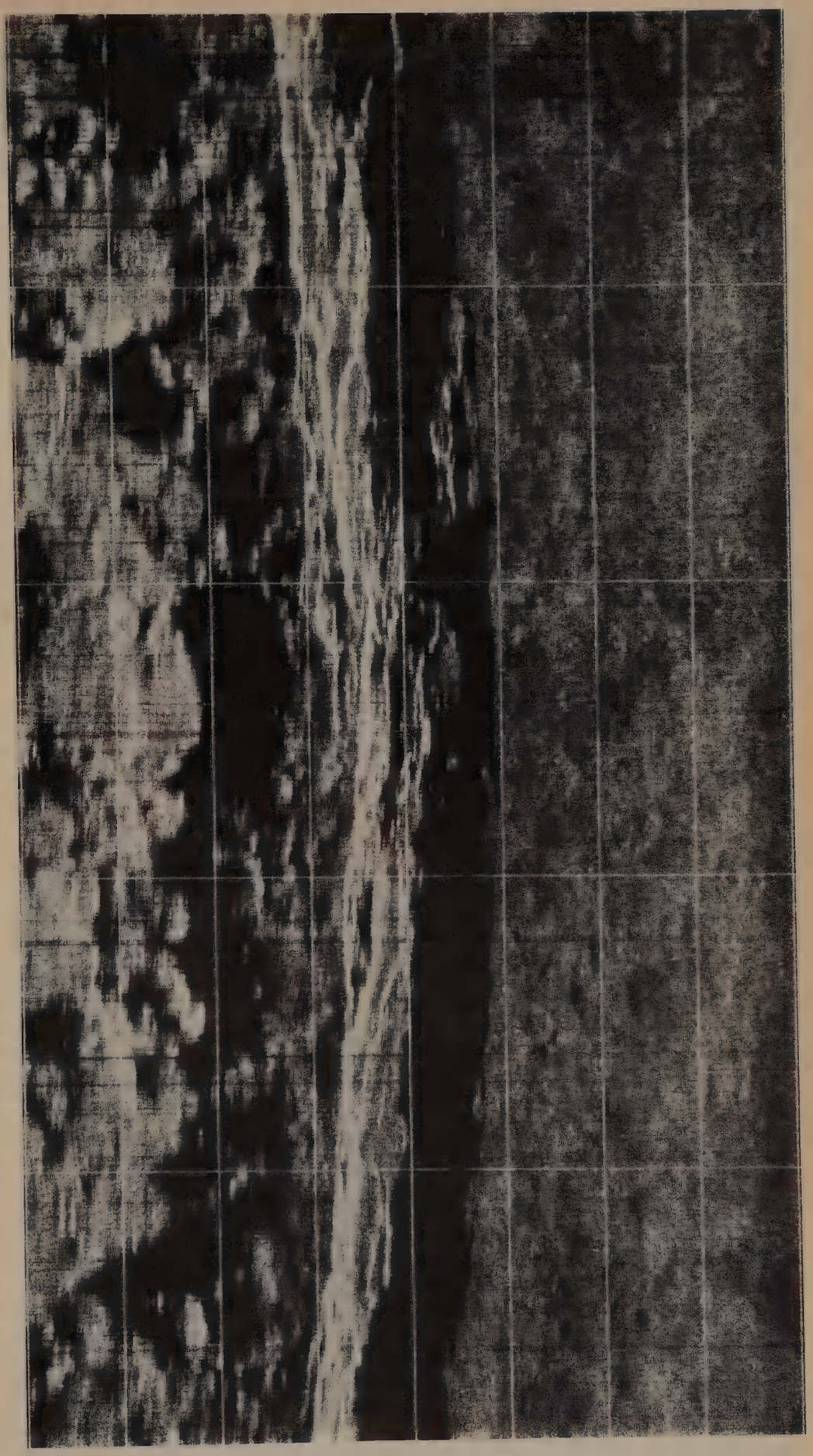
Because the Mid-Atlantic Ridge and the East Pacific Rise are almost invisible to instruments that measure the gravity field, other techniques must be employed to map them. The most significant are methods based on recording acoustic pulses with a very high frequency projected from a device on shipboard to the ocean floor. The depth under the ship can be calculated from the travel time of the pulse from the ship to the bottom and back. In addition to the capacity to map spreading centers such sonic methods can provide much greater detail than a satellite can.

Sonic techniques are not new, but recent innovations have made them much more valuable to earth scientists than they were previously. The most significant innovation is arrays of sonar sources and receivers mounted in a single device. The information from the multiple beams can quickly be combined into a detailed topographic picture. An even more recent innovation is to mount such arrays on vehicles towed at considerable depth behind a ship.

The system called SeaBeam, made by the General Instrument Corporation, has an array of 16 beams, each with a frequency of 12,000 hertz (cycles per second). Each beam is a little less than three degrees wide. The beams are projected out in the shape of a fan along a line perpendicular to the ship's course, enabling the device to map a swath of sea floor two-thirds the water depth as the ship advances.

High-resolution SeaBeam mapping is a powerful tool for examining the sea floor, particularly the parts of the ridge system where the plates diverge rapidly. The spreading rate at the ridge crest can vary considerably. The Mid-Atlantic Ridge is a medium-to-slow spreading center with a rate of about five centimeters per year. The East Pacific Rise, on the other hand, is a fast spreading center where the rate of divergence can be as high as 15 centimeters per year. At fast spreading centers the features at the crest of the ridge tend to be concentrated in a fairly narrow zone that can be taken in with a single pass of SeaBeam; hence the system is particularly useful for mapping areas such as the crest of the East Pacific Rise.

The axis of the East Pacific Rise runs roughly north-south off the coast of Central America and South America. At its crest is a ridge about 30 kilometers across and 500 meters high. The axial ridge has recently been the subject of intensive work both by earth scientists interested in the fast spreading center and mining companies interested in the sulfide compounds deposited along the



AXIAL RIFT delineates the exact center of the midocean ridge: it is the line along which lava flows out onto the surface. The rift shown is on the East Pacific Rise, a major spreading center that runs along the coast of Mexico, Central America and South America. The image was made with the sonar system SeaMarc I. A source of high-frequency acoustic pulses is housed in a "fish" that is towed close to the ocean floor behind a ship. The pulse is projected to the side and the intensity of the reflected beam is recorded. The data are later converted into an image of the ocean bottom. In this sonar image highly reflecting surfaces are light and nonreflecting surfaces and shadows are dark. The image was made during a cruise over the Clipperton Fracture Zone at 21 degrees north in the Pacific led by William B. F. Ryan of Lamont-Doherty.



BLACK SMOKERS are vents through which hot water is expelled from the crust. The particles that give the plume its dark color are sulfides that have been dissolved out of the crustal rock. Near the midocean ridge seawater penetrates cracks in the newly formed crust. The water is heated and propelled upward through the vents. The photographs on this page were taken on the East Pacific Rise from the manned submersible *Cyana*. This one was taken during an expedition led by Roger Hékinian of the Centre National pour l'Exploitation des Océans.



FISSURE CUTS THROUGH THE SURFACE of the oceanic crust near the crest of the East Pacific Rise. Such fissures are caused by plate motion, which generates lateral stresses, and also by the contraction of the crust. The fissures nearest the ridge crest are active sources of lava flow and are often concealed by the flows. Those farther from the ridge crest, like the one shown, no longer expel lava. This photograph was taken in 1978 during the first exploration of the East Pacific Rise by manned submersible in an expedition led by the author.

ridge by the heated seawater rising from the vents.

The crest of the ridge includes a zone of active volcanism. In the course of dives with the French submersible *Cyana* in early 1982 other investigators and I observed that the area of current volcanism is only one to two kilometers wide, so that it can easily be mapped in one SeaBeam pass. Indeed, about two years earlier a SeaBeam map of the "Great Pacific Highway," the narrow, flat crest of the rise, was made from the French research vessel *Jean Charcot*. Similar maps have since been made by other vessels, and the detailed topography of the East Pacific Rise is becoming clear.

The simultaneous employment of sonar mapping and manned submersibles on the East Pacific Rise has revealed a dramatic topographic pattern that could modify the notion of how the spreading center works. In the northern part of the East Pacific Rise major fracture zones marked by transform faults interrupt the ridge crest every 200 to 300 kilometers. Between the large faults are many smaller faults, some as little as 10 kilometers apart.

Near the transform fault the ridge is quite deep, and it rises to a peak between each pair of faults. The topographic high is usually equidistant from the two faults; the overall swell is about 500 meters over the 200-to-300-kilometer distance. Thus in a profile taken along the strike the ridge has the appearance of a gently rising hill. Since the profile of the ridge transverse to the strike is also that of a gentle rise, the area between a pair of transform faults is shaped like a low dome.

A second set of smaller domes protrudes from the large structure like a row of blisters along the ridge crest. This second set of domes was discovered in multibeam sonar studies by Robert D. Ballard of the Woods Hole Oceanographic Institution, Roger Hékinian of the Centre National pour l'Exploitation des Océans and me on the East Pacific Rise between the Orozco Fracture Zone at 15 degrees north and the Clipperton Fracture Zone at 10 degrees north. The information from the sonar maps was augmented by dives in *Cyana*.

The small domes are bounded by the faults that cut the ridge between transform faults. The small projections are about 100 meters high. It is probable that each dome corresponds to a spreading center: a small area where the manufacture of new crust goes on independent of the adjacent ridge segments. Thus instead of a single great factory the midocean ridge could be a string of small adjacent workshops. Hans Schouten and his colleagues at Woods Hole have suggested that the midocean ridge is indeed such a chain of adjacent

spreading cells separated by fracture zones; the cells could be stable for long periods of time. Schouten's hypothesis implies that oceanic crust is not created as a single homogeneous mass but is made in long narrow ribbons laid side by side with fracture zones in between.

Although SeaBeam is the most convenient sonar system for mapping along-strike topography, two other sonar systems can give a more detailed picture of the ocean floor than SeaBeam can. Both are "side-looking" sonars, meaning that the acoustic pulses are projected to the side from devices towed behind the ship near the ocean floor. The Geological Long Range Inclined ASDIC (GLORIA) was developed at the Institute of Oceanographic Sciences at Wormley in England. The sonar transmitters encased in the "fish," a neutrally buoyant housing, emit pulses of sound with a frequency of 6,200 to 6,800 hertz. GLORIA can map a swath of sea floor about 30 kilometers wide.

A side-looking sonar system has one great advantage over a single-channel sonic profiling system operated from shipboard. In the shipboard systems vertical features such as scarps are difficult to detect because they yield little upward reflection of the acoustic beams. In the side-looking system, however, vertical features show up clearly because they offer excellent surfaces for horizontal reflection. Thus at the midocean ridges GLORIA has made plain the pattern of inward-facing scarps along the faults that parallel the ridge axis.

Each side-looking sonar system has its own frequency and arrangement of beams, hence it is best suited for a particular purpose. SeaMarc I was developed by International Submarine Technology, Ltd., to search for the remains of the liner *Titanic*. The SeaMarc device is towed in a fish from 100 to 400 meters above the sea floor. It transmits two roughly horizontal beams with a frequency of 27,000 to 30,000 hertz. SeaMarc can map a section of sea floor about five kilometers wide; its scale is thus between that of SeaBeam and that of GLORIA. Combining results from the SeaBeam, GLORIA and SeaMarc systems is beginning to provide a rich picture of the spreading centers. On a recent cruise led by William B. F. Ryan of Lamont-Doherty and J. Paul Fox of the University of Rhode Island the SeaMarc system was employed to map the East Pacific Rise between the Clipperton Fracture Zone and the Orozco Fracture Zone. The structures associated with the creation of oceanic crust could be seen as clearly as if they had been observed from the air.

In the past five years satellite data and sonar recordings have filled more detail in the map of the ocean floor than has



The 757 is the most fuel-efficient jetliner in the sky. In passenger comfort, it's superior to any other plane its size. Here's the perfect match of advanced technology and passenger conveniences. It's one way Boeing is helping to keep air fares one of the world's best travel values.

BOEING
Getting people together.



SOME COFFEES YOU SAMPLE. OURS, YOU LOVE.

That sample costs only $10.75, a remarkable price for a selection of three one-pound packages of one of the world's finest gourmet blends and roasts, Cap Saurage's Community Coffee—a taste that has led three generations of customers to savor and enjoy particularly flavorful lives. It is now available at selected gourmet food stores.

If you wish, we will send you this selection along with our gourmet coffee book and the name of your nearest retailer, so you may experience the finest blends of Brazilian Santos, Colombian Medellin-Armenia and Mexican Altura coffees: an introduction to the pleasures of different roasts from a subtle light medium to the peak flavor in our dark roasted coffees.

Call Toll Free 800-535-9901 (with your credit card ready) seven days a week, 7 a.m.–7 p.m. Or, send $10.75 with your name and address to Community Coffee Company, P.O. Box. 3778, Dept. SC1, Baton Rouge, Louisiana 70821-3778. Specify choice of whole bean, regular or filter grind.

Community Coffee
BRAND®



Passengers rate the Boeing 767 as a superb flying experience. It has two wide aisles and is the only wide-body with seven-abreast seating, so nearly everyone has either a window or aisle seat. Pilots rate the 21st Century flight control system as the most advanced in the world. BOEING
Getting people together.



RECENT PUBLICATIONS IN NATURAL HISTORY

With the flood of scientific literature being published, how can scientists and libraries keep up-to-date? **Recent Publications in Natural History**, a quarterly bibliography published by the Department of Library Services of the American Museum of Natural History, can help. Each issue lists new publications in zoology, anthropology, earth sciences, museology, history of science and related subjects.

Each issue includes

BOOK REVIEWS: full-length scholarly reviews by leading scientists
WORLDWIDE COVERAGE: publications in any language from any country
24 SUBJECT HEADINGS: with cross-references
FULL BIBLIOGRAPHIC CITATIONS WITH ANNOTATIONS: including ISBN or ISSN, price when available and addresses of smaller or lesser-known publishers

SUBSCRIPTION INFORMATION
$10.00 per year (4 issues), by volume only. Outside U.S. and Canada, add $3.00 per year for each subscription. Subscribe to Volume 2 (1984) and receive a **FREE** subscription for Volume 1 (1983).

American Museum of Natural History
RPINH Subscriptions (Library)
Central Park West at 79th Street
New York, NY 10024 USA

been added in any comparable period before. As a result earth scientists now have almost as clear a picture of some parts of the midocean ridge as they do of structures on land. To penetrate below the crust and develop a picture of its composition and structure, however, other techniques are needed.

Much of the current understanding of the layers of the oceanic crust comes from recordings of seismic waves, both those generated naturally by earthquakes and those generated experimentally by explosions or special air guns. Indeed, the definition of the crust was originally formulated on the basis of work with seismic waves. The Mohorovičić discontinuity, or Moho, which divides the crust from the mantle, was first detected because of its capacity to reflect seismic waves.

The speed with which seismic waves travel depends on the temperature, pressure and composition of the rock medium. Hence recordings made at a distance from an earthquake or an experimental explosion can yield clues to the makeup of the intervening material. Two main types of waves are studied in the seismic work: body waves, which tend to travel through a particular layer, and surface waves, which tend to travel along the boundary between two layers.

Both body waves and surface waves can consist of two other types of waves: *P* waves and *S* waves. *P* waves are analogous to sound waves in air. As the *P* wave passes, the rock is compressed and expanded only in the direction of wave motion. *S* waves subject the rock to shear stresses perpendicular to the direction of wave motion. *S* waves can travel only through solids; *P* waves can travel through solids, liquids and gases.

In the late 1950's Russell W. Raitt of the Scripps Institution of Oceanography proposed a seismic model of the crust. In Raitt's model the crust was composed of three layers that could be distinguished on the basis of the *P*-wave velocity. Layer 1 consisted of the sedimentary cover, Layer 2 the oceanic basement and Layer 3 the oceanic layer. The underlying mantle was designated Layer 4. The *P*-wave velocity was held to be constant in each layer.

Raitt's hypothesis included three simplifying assumptions. The first was that the wave velocity in the rock always increases with depth. In general this is the case, because the compression of the rock increases its "resonance" as a medium for wave motion. There are, however, exceptions to the rule, and the velocity pattern in the crust is complex.

Raitt's second assumption was that the boundary between any two layers of the crust is a horizontal plane. His third assumption was that the layers are quite thick in relation to the character-

istic length of the waves. In general a seismic wave can give information only about features that are considerably larger than the length of the wave. Features that are smaller than the wavelength have little effect on the path of the wave and therefore cannot be detected when the wave motion is analyzed. Experimental explosions (which were the main source of seismic waves in geologic work in the 1950's) generally give rise to waves with a wavelength of from .5 kilometer to two kilometers; thus they yield information only about components of the crust that are a few kilometers thick or thicker.

Recent work has shown that all three assumptions are drastic oversimplifications. Even the idea that the velocity in the layers is constant has now been called into question. Paul Spudich and John A. Orcutt of Scripps and G. Michael Purdy of Woods Hole have put forth a model of a crust in which layers are defined not by the absolute wave velocity but by the velocity gradient: the change in wave velocity with depth. Recent seismic work has been much more concerned with Layer 2 and Layer 3 than with Layer 1, where sediment often gives inconsistent seismic results.

According to Spudich and Orcutt, in Layer 2 the velocity gradient is quite steep. The wave velocity increases by from one to two kilometers per second with each kilometer of depth in the crust. In mathematical notation such a gradient is written as 1 to 2 s^{-1}. In Layer 3 the gradient is about .1 s^{-1}. Thus the *P*-wave velocity in Layer 3 is almost uniform, which makes Layer 3 the best defined region of the crust from the seismic point of view. In Layer 4, the upper mantle, the wave velocity exceeds 7.8 kilometers per second.

Some seismologists argue that there is an additional low-velocity zone between Layer 3 and Layer 4 corresponding to the transition from the crust to the mantle, but the evidence for such a layer is scanty. It is known that the transition from the crust to the mantle takes place over a distance of three or four kilometers. Waves with a wavelength short enough to give a high-resolution picture of the Moho, however, become quite attenuated in traveling to the bottom of the crust and back to the surface, and therefore little is known in detail about the transition layer.

Purdy took a seismic profile southwest of Bermuda in oceanic crust created during the Mesozoic era about 140 million years ago. Carefully allowing for the effects of a variable sediment cover, he found that the crust was 7.2 kilometers thick above a 500-meter transition zone to the mantle. The *P*-wave velocity at the top of the crust was five kilometers per second. Layer 2, defined as the



World travelers choose the 737 because it has one of the best on-time records. The 737-300 is the latest addition to the family. This new jetliner, with added passenger amenities, advanced flight controls and quieter engines, will make flying an even better travel experience.

BOEING
Getting people together.

A BERLITZ® LANGUAGE COURSE
in a handsome briefcase
for only $125
if *you* pay for it
BUT
. . .wouldn't your *company*
be interested in having you
speak a second language?



These days a second language can help you help your company and so help *you* earn more. The many hispanic nationals now in our work force respond more readily to someone who speaks their language. Also, selling today is international. Whatever language it would pay you to speak, Berlitz has a do-it-yourself course in a handsome briefcase you can take anywhere. Contents—90 minute introductory tape, 40 lessons on five 60-minute cassettes, 6 illustrated word and phrase books, rotary verb finder to help you find verb tenses instantly. If not satisfied, return within 10 days and get your money back. Learn from the leader!

For your convenience on credit card orders dial toll-free
1-800-228-2028 ext. 35
and **refer to Dept. 3165**
24 hours a day,
7 days a week.

Why not give us a ring—right now! (our expense)

(Nebraska residents dial 402-571-4900)

Berlitz Publications, Inc., Dept. 3165
3490 Lawson Blvd., Oceanside, NY 11572

Send **Berlitz Comprehensive Cassette Course(s)** checked. $125 each in briefcase plus $4 for shipping and insured delivery.

☐ **French** 86100 ☐ **German** 86101 ☐ **Italian** 86102 ☐ **Spanish** 86103

Enclosed check _____ money order _____ payable to Berlitz.
Or charge my AMEX _____ Diners Club _____ VISA _____
MasterCard Interbank # _____ Exp. Date _____
Card # _____ Exp. Date _____
Name _____
Address _____
City _____ State _____ Zip _____

N.Y. residents add sales tax. Allow 4 weeks for delivery.

region where the gradient was greater than .64 s^{-1}, was 2.3 kilometers thick. Layer 3, with a gradient of .1 s^{-1} or less, was 4.9 kilometers thick.

In the crust where Purdy did his work Layer 3 is subdivided into two levels: an upper level 1.7 kilometers thick with a gradient of .1 s^{-1} and a lower level 3.2 kilometers thick where the waves show no change in velocity with depth. The velocity in the lower level of Layer 3 is seven kilometers per second. Thus the oceanic layer (Layer 3), seismically defined, is about five kilometers thick and the basement layer above it (Layer 2) is half as thick.

Several workers have attempted to go beyond the structural profile given by the velocity gradients to infer the composition of the rock layers. This is generally done by comparing the *P*-wave velocity with the *S*-wave velocity for rock at a given depth in the crust. Different kinds of rock have characteristic ratios of *P*- to *S*-wave velocities and hence some deductions about the rock type can be made on the basis of the seismic data. Unfortunately for the seismologist the wave-velocity ratios for the rock types are not unique. The velocity change from Layer 2 to Layer 3 could therefore be interpreted as either a change in metamorphic rock type (from greenschist above to amphibolite facies below) or a change in lithologic type (from basalt above to gabbro below).

It is possible that knowledge of the exact composition of the lower layers of the oceanic crust will have to wait until drills penetrate deeper in the crust than they have done so far. In the winter of 1981 a drill on the *Glomar Challenger* in the six-nation International Program for Ocean Drilling succeeded for the first time in going deeper than one kilometer. The drill hole, 1,076 meters deep and designated 504-*B*, is in the Costa Rica rift, which crosses the East Pacific Rise between the Galápagos Islands and South America.

Although a one-kilometer hole seems modest compared with the full seven-kilometer depth of the crust, the achievement is a considerable one because of the problems of drilling in the ocean. The difficulties encountered in drilling holes in the oceanic crust have included the twisting and breaking of the drill string and severe wear on the bits. The drilling of Hole 504-*B* was relatively free of such problems. What does Hole 504-*B* tell us about the crust?

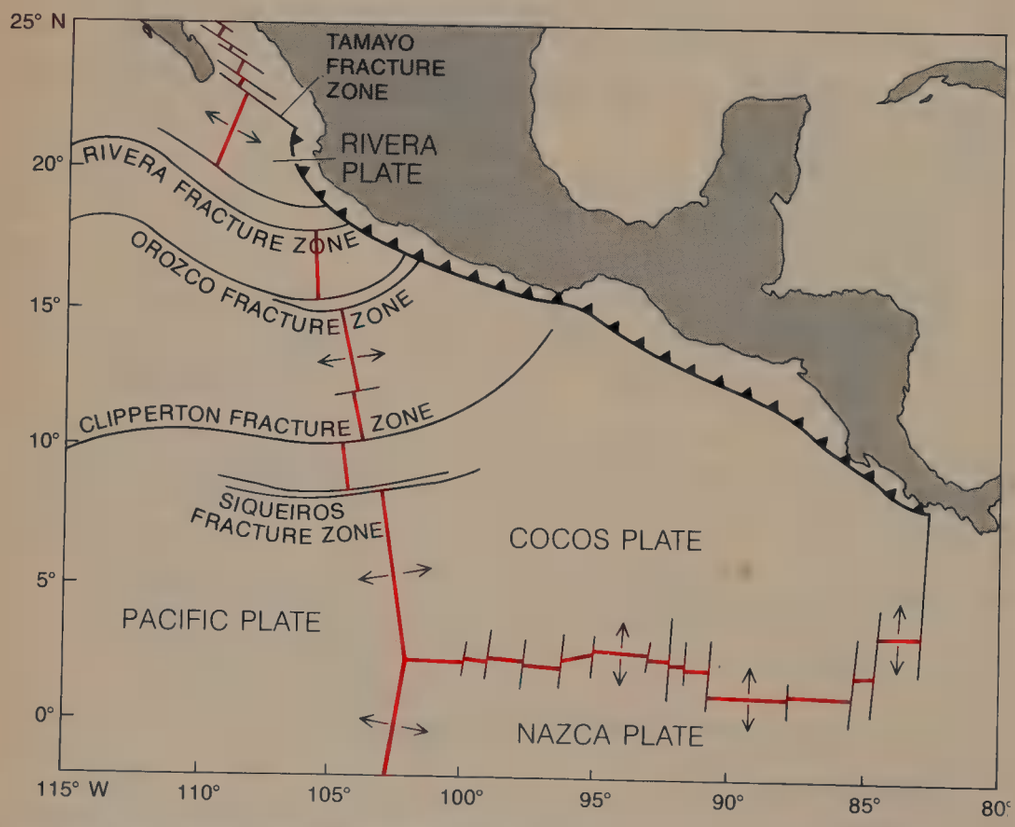
The crust at the drill hole is six million years old. Because the Pacific at the latitude of the Costa Rica rift is warm and biologically active, the surface of the crust is already covered with a layer of sediment 275 meters thick. Below the sediments the upper 575 meters of the basement consists of pillow lavas and the rocks called breccias and hyaloclastites, which are produced by the melding of small pieces of fractured basalt into a single mass under pressure. Between 575 and 780 meters the first dikes appear and extensive breccias are interspersed with a few pillow lavas. From 780 meters to the bottom of the hole are massive basalts, abundant dikes and a notable lack of either pillow lavas or fractured material.

Recordings of the in situ pressure deep down in Hole 504-*B* made by Roger N. Anderson of Lamont-Doherty and Mark Zoback of the U.S. Geological Survey are clarifying the circulation of water through the upper layers of the crust. Anderson and Zoback employed an inflatable device called a packer to isolate a section of the hole. The pressure in the isolated zone was measured from a "go devil" developed by the oil industry to explore oil and gas wells. Intriguingly, Anderson and Zoback observed that the water halfway down the hole is under less pressure than the water at the top. The difference in pressure is about eight bars. (One bar is 14.7 pounds per square inch.)

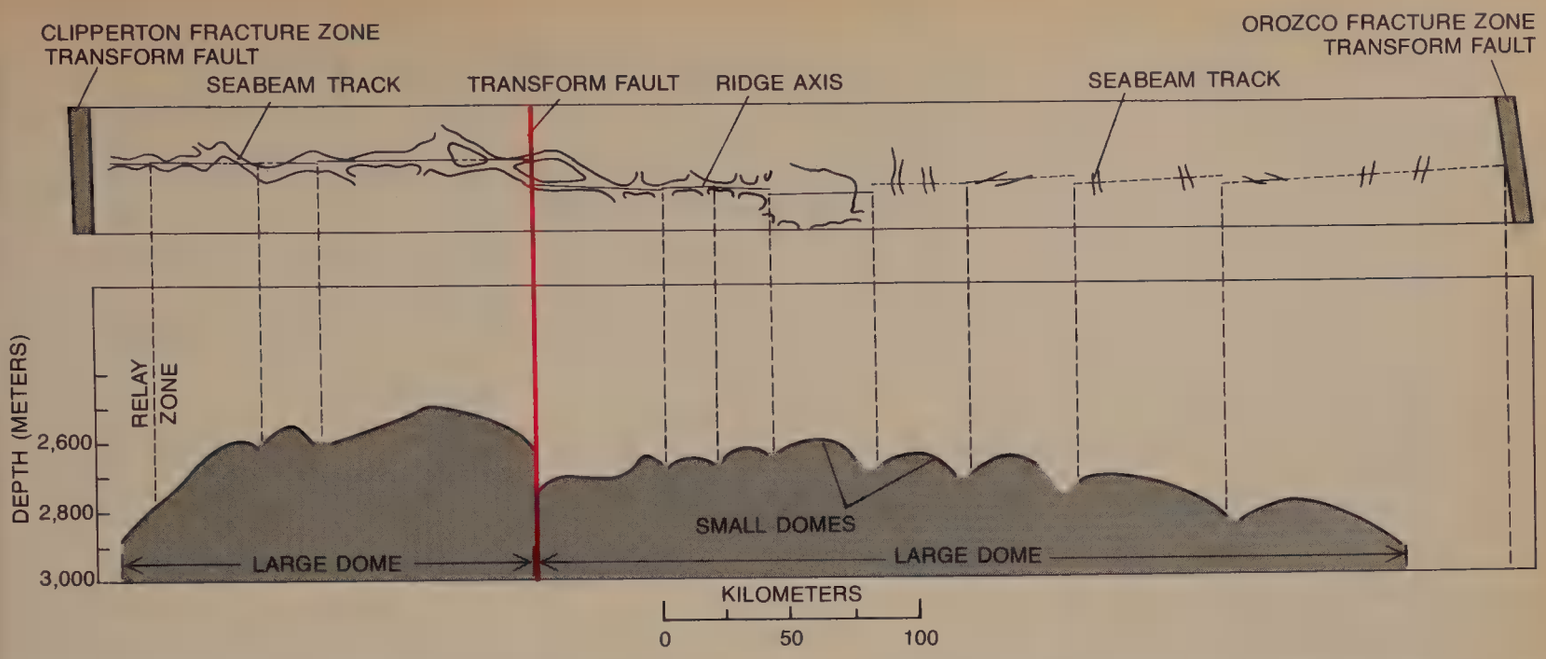
The mechanism of the remarkable reversal of the pressure gradient is not yet understood. The underpressure in the hole could, however, be due to a heat-convection cycle in the mantle. The relatively low pressure in the lower layers of the crust could help to explain the hydrothermal circulation; it would tend to pull seawater from the ocean floor down into the cracks in the crust.

The depth of the hydrothermal circulation is currently a matter of considerable controversy. It has been suggested that the circulating water penetrates deeply enough to have a role in regulating the operation of the magma chamber under the midocean-ridge axis. If enough seawater reaches the lower layers of the crust, it could cool the magma and cause it to solidify. The lava would therefore stop flowing onto the surface of the ridge until the plates pulled apart enough to renew the decompression of the mantle rock.

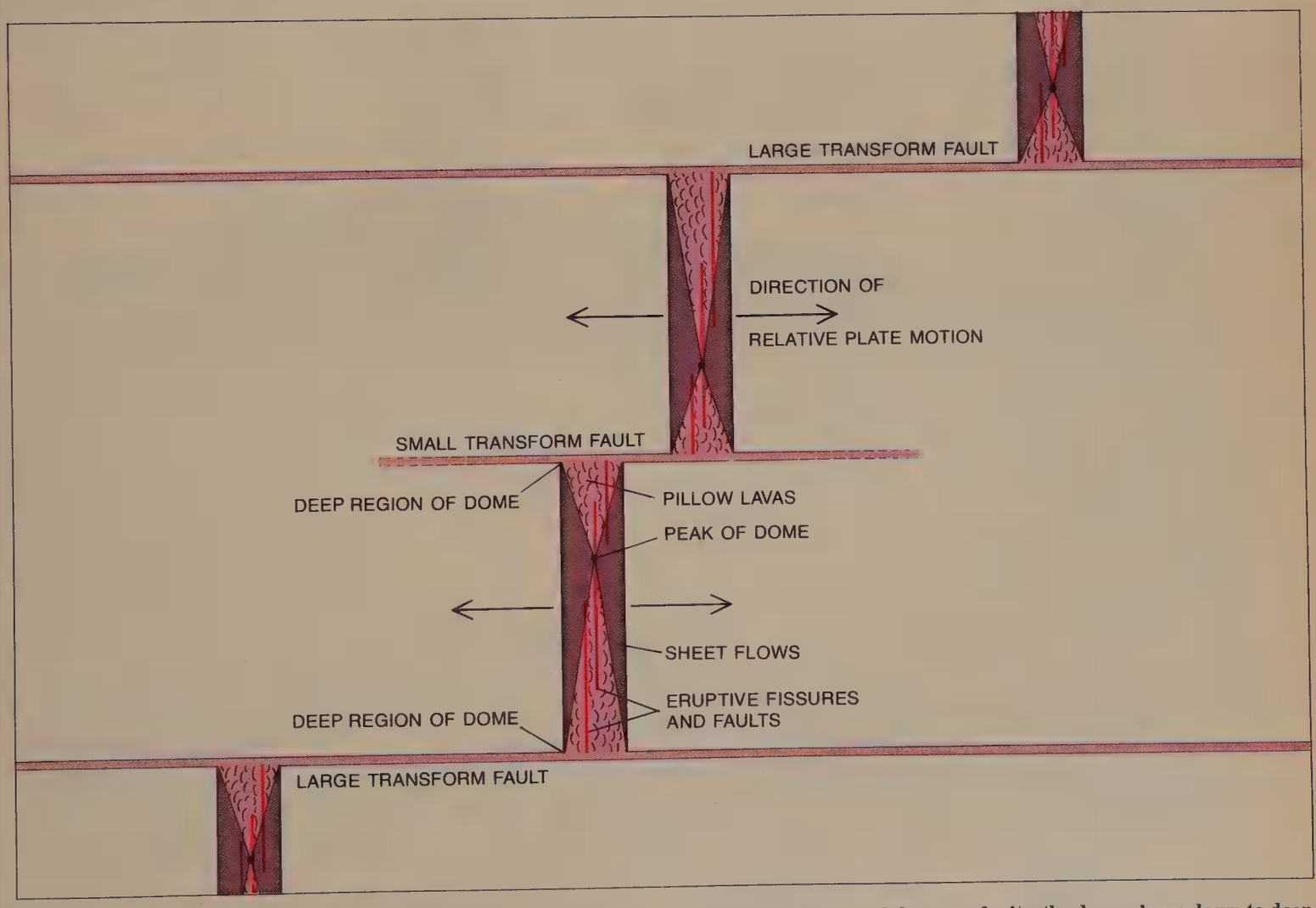
For the circulating water to have such an effect it would have to penetrate into the gabbro in Layer 3. Hole 504-*B* did not go deep enough to settle the question, but it is significant that the rocks recovered from a depth of 600 meters on down to the bottom of the hole show a pattern of alteration that is quite compatible with the fluxing of heated seawater through them. Moreover, the rocks from the deepest part of the hole are among the most extensively altered. Such alteration would have had to take



EAST PACIFIC RISE OFF SOUTHERN MEXICO has a complex geometry that includes many faults across the strike, or main axis, of the ridge. The axis of the rise runs along the boundary between the Pacific plate and the Cocos plate; arrows indicate the direction of plate motion. Fracture zones mark the points where major transform faults cross the strike. At the fracture zone the ridge axis is offset. Within the offset the direction of movement of the crust on one side of the fault is opposite to that of the crust on the other side. Between the fracture zones are many smaller faults. Much of the author's recent work in the *Cyana* and the research vessel *Jean Charcot* has been done between the Clipperton and Orozco fracture zones.



SUPERPOSED DOMES at the crest of the East Pacific Rise could give clues to how the spreading center at the midocean ridge operates. The upper panel shows part of the East Pacific Rise in plan view as mapped by the SeaBeam sonar system. The mapping was done by the author with Hékinian and Robert D. Ballard of the Woods Hole Oceanographic Institution. The ridge axis is marked by the central line; broken segments stand for regions where the axis had to be guessed at. The axis is marked by active volcanic fissures. Pairs of parallel lines show areas that have been mapped by the SeaBeam sonar system. Ridge offsets mark transform faults and relay zones, where small faults cross the strike. The lower panel shows the same ridge segment in profile. Two large domes run between transform faults, with the peak about halfway between the faults. A series of small domes bounded by relay zones protrude from each large dome.



HYPOTHESIS proposed by the author and Ballard to explain how the spreading center operates includes the idea that each large dome between transform faults is an individual spreading cell: an area where the creation of oceanic crust proceeds independently. In this schematic diagram two complete domes are shown. The broken line of eruptive fissures from which lava flows marks the ridge crest. Each dome has a peak between faults; the dome slopes down to deep regions next to the fault. Near the peak the ridge crest is covered mainly by sheet flows. Farther down, the slope is covered by pillow lavas. Which form of lava is predominant depends on the position on the dome. If each dome is a spreading cell, the crust could be created in thin ribbons extending from the sides of the spreading cell.

place before the sediments were deposited, capping the basement and preventing the flow of water into it.

The clues to the operation of the magma chamber gained from Hole 504-*B* are being supplemented by seismic work. In investigating the magma chamber both the reflection and refraction of seismic waves are observed. In the reflection experiments air guns towed behind a ship are generally employed as the energy source. The waves from the air guns pass down to the crust and are reflected upward to the ship, where their travel time and amplitude are recorded. In the refraction experiments the energy propagates at the interface between rock layers and can be recorded some distance away on the sea floor with an ocean-bottom seismometer or at the sea surface. By combining reflection and refraction results the speed of the waves through the structures of the crust can be calculated.

The speed of seismic waves is reduced considerably by rock in the molten state; therefore a low-velocity zone in the crust could correspond to a magma chamber. There is an area under the East Pacific Rise where the refracted waves are attenuated or much slowed; the area is probably a crustal magma chamber. Magma is an efficient reflector of seismic energy, so that in the reflection work the top of the magma chamber comes out as a strong, roughly flat reflecting surface two or three kilometers below the sea floor. In the spreading center on the East Pacific Rise at nine degrees north and also in the Lau Basin near the Fiji Islands the reflection at the top of the chamber is about four kilometers wide.

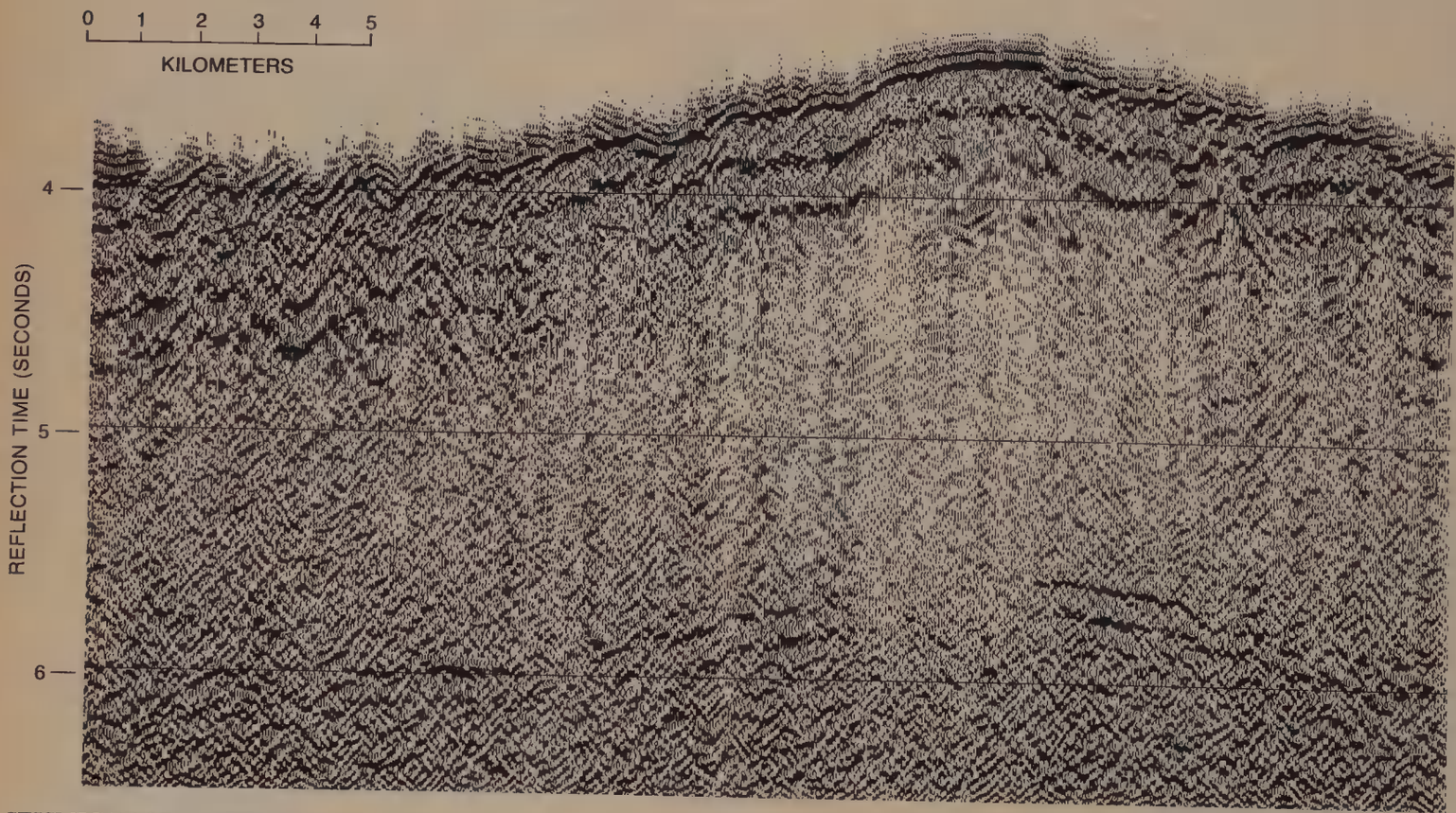
Refraction experiments done by Brian T. R. Lewis and his colleagues at the University of Washington also make it clear that the magma chamber is quite narrow. Lewis believes it is even less than four kilometers wide. Thus the entire oceanic crust is created by a thin tube of molten rock running a few kilometers below the crest of the midocean ridge.

The "root" of the midocean ridge, the underlying structures in the mantle, is also being seismically probed. One means of examining the deeper layers is the surface waves called Rayleigh waves. Donald W. Forsyth of Brown University, Nicole Girardin of the Institut de Physique du Globe of the University of Paris and Wolfgang Jacoby of the University of Frankfurt studied the passage of Rayleigh waves through young lithosphere in the Pacific and also along the Reykjanes Ridge south of Iceland. They found that the low *S*-wave velocity extends down to 60 kilometers.

Additional data were supplied by waves from a strong earthquake in Uzbekistan in the southern U.S.S.R. in May, 1976. Rayleigh waves from the earthquake with a period of 300 to 400 seconds, which can make several circuits of the earth, were recorded with ultralong-wavelength seismometers at Los Angeles and at the South Pole. The passage of the long-wavelength waves shows that the low-velocity zone under the East Pacific Rise extends to a depth of 200 kilometers, with the minimum velocity at about 100 kilometers. Thus the midocean ridge has deep roots.

Stuart A. Hall of the University of Houston and his colleagues have advanced a quantitative model of the magma chamber and its root in the mantle. Their work is intended to account for the small gravitational anomalies on the crest of the midocean ridges. As I mentioned above, the ridge is approximately in isostatic equilibrium and so is not associated with large anomalies in the gravitational field. There are, however, small anomalies directly over the ridge crest. At the Mid-Atlantic Ridge, where the spreading rate is low, there is a negative anomaly; at the East Pacific Rise, where the spreading rate is high, there is a small positive anomaly.

According to Hall and his colleagues,



SEISMIC PROFILE offers evidence that the magma chamber under the midocean ridge is quite narrow. The profile was made by Peter Buhl of Lamont-Doherty at nine degrees 30 minutes north on the East Pacific Rise. The peak on the upper border of the profile is the ridge crest. The dark line at a reflection time of about six seconds is the Mohorovičić discontinuity, or Moho: the boundary between the crust and the mantle. A seismic profile is made by generating strong acoustic waves and recording their reflections from various levels of the crust. The travel times and amplitudes of the reflected seismic waves yield information about the character of the crustal rock. In a partially liquid magma the seismic waves are considerably retarded. The break in the Moho directly under the ridge crest indicates the presence of magma. The width of the break suggests that at the level of the Moho the magma chamber is less than two kilometers across.

the gravitational anomalies could both be explained by one type of magma chamber. They hypothesize that the density of the material in the magma chamber is about 2.75 grams per cubic centimeter, or 1 percent lower than that of the surrounding rock. The mantle root that is made up of gabbro has a density of three grams per cubic centimeter, or 6 percent lower than that of the surrounding mantle rock. The density of the chamber and the mantle root is thus quite close to the density of the adjacent rock.

Hall and his co-workers conclude that the small gravitational anomalies over the ridges are therefore due not to variations in density but to topographic features at the ridge crest. On the slow-spreading Mid-Atlantic Ridge there is a rift valley along the ridge axis. On the fast-spreading East Pacific Rise the axis is marked by a ridge.

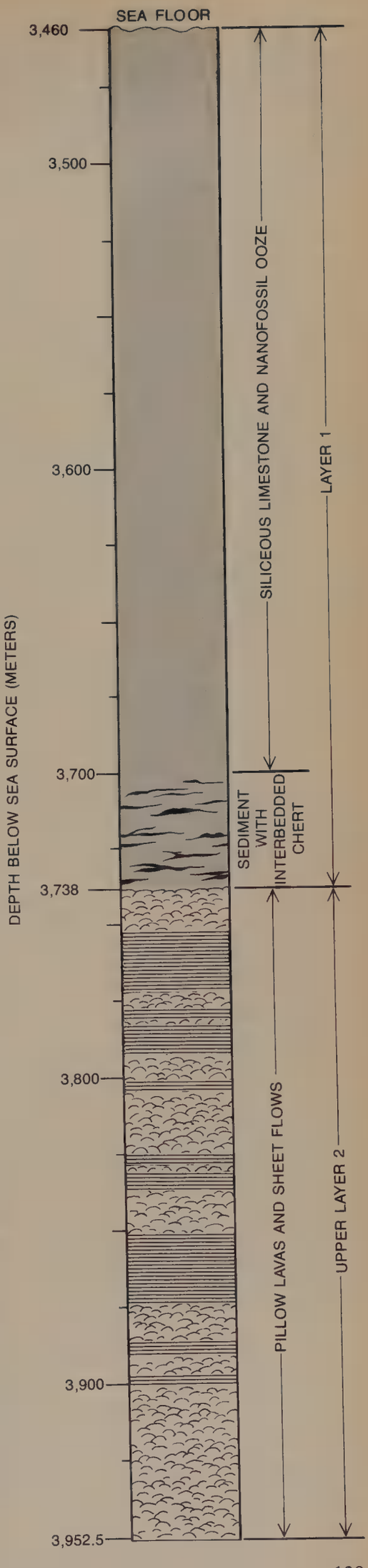
The newest source of information about the structure of the oceanic crust is not seismic waves or gravity data but electromagnetic radiation. It is now possible to measure undersea electric and magnetic fields with considerable accuracy by means of receivers on the ocean bottom. If a source of electric current is put on the sea floor some distance away from such a receiver, the electric field induced by the current travels through the rock. The measured intensity of the electromagnetic indicates how well the intervening section of the crust conducts electricity.

The electrical conductivity of rock is affected by chemical composition, temperature and the extent of melting. A deep electrical conductivity log can therefore be quite informative about the rock of the crust and the upper mantle. As a source of electromagnetic energy Charles S. Cox and Peter Young of Scripps have employed an insulated wire 800 meters long with bared ends. The wire acts as a horizontal electric dipole, with the return flow of current being through the ocean. The wire is laid on the sea floor at the end of a cable attached to a ship. An alternating current with a peak of about 70 amperes is passed through the wire. The energy transmitted by the dipole has a frequency of about one hertz. A pair of cruciform antennas with arms nine meters long are put on the sea floor 19 kilometers away from the transmitter. With this setup Cox and Young have recorded electric field signals as strong as 10^{-10} volt per meter. Since the "noise," or background electric field on the sea floor, is 10^{-12} volt per meter, 100 times weaker than the recorded signal, the finding is significant.

The pattern of received signals shows that there are two layers in the crust with quite different electrical conductivity. The upper layer is at most 1.5 kilometers thick and has a fairly high electrical conductivity: about .1 Siemens per meter. The upper layer corresponds to the relatively young fractured basalts found near the top of Hole 504-*B*; the measured conductivity in the two places is about equal. In both places the presence of seawater that has penetrated the fractures greatly increases the electrical conductivity.

Below the conductive layer is a region, extending down to about six or seven kilometers, in which the conductivity is much lower: about .004 Siemens per meter. The measured conductivity in the lower region is probably an average for the lower crust and upper mantle. In the deep parts of the crust the conductivity is due to seawater penetrating the gabbro layer. In the mantle the conductivity results from the passage of the electric signal through minerals in the hot rock.

Electromagnetic measurements can yield unique information about the changes in temperature with depth and the presence of molten rock in the deep regions of the crust. Therefore this new method could in the next few years throw much light on the zone near the Moho, which is not easily investigated with other techniques. Novel and intriguing as it is, electromagnetic observation is only one of an array of methods that are currently being harnessed to probe the upper part of the earth under the ocean. Theories about the oceanic crust advance, even more than theories in other areas of science, only in conjunction with new methods of observation. The past decade has been a period of remarkably rapid development of such methods. Given the inevitable lag between observation and the formulation of new theories, in the next decade there could be developed a new and more accurate picture of the thin layer that covers most of the surface of the earth.



DEEPEST HOLE yet drilled in the oceanic crust is Hole 504-*B* on the Costa Rica rift between the Galápagos Islands and South America. Drill cores from Hole 504-*B*, such as the one shown in schematic form, are providing new information about the structure of the crust. The depth below the sea surface is given at the left; the core depicted extends for about half the full depth of the hole. The crust on the Costa Rica rift is about six million years old. It is covered by a sediment layer 275 meters thick, made up mainly of the remains of microscopic marine plants and animals. Below 275 meters is the oceanic basement, composed of pillow lavas and sheet flows. Intriguingly, it has been shown that the in situ pressure at the bottom of the core in the illustration is less than the pressure at the ocean floor. The difference in pressure could pull seawater down into cracks in the crust.

The Continental Crust

It is much older than oceanic crust, some of it dating back nearly four billion years. It is constantly reworked, however, by cycles of tectonics, volcanism, erosion and sedimentation

by B. Clark Burchfiel

For a little more than 200 years earth scientists have been studying the geology of the continents, hoping to reconstruct a record of the history of the earth. It is a daunting project. The crust underlying the oceans is rapidly created; it remains intact and relatively undeformed for most of its short lifetime, and then it is rapidly destroyed. The oldest crust in the earth's ocean basins today is less than 200 million years old. In contrast, the crust making up the continents is created and modified by a variety of physical and chemical processes, often undergoing many phases of deformation and reworking that produce a complex worldwide pattern in which belts of deformed rock hundreds of kilometers wide and thousands of kilometers long are invaded by intrusions of igneous rock and locally overlain by a thin veneer of younger sedimentary rock. In addition continental crust largely resists the processes that destroy oceanic crust. The most ancient parts of the continents are about 3.8 billion years old. Thus the continental crust holds a complex and fragmentary record of the evolutionary and dynamic processes operating through 85 percent of the earth's 4.6-billion-year history.

Continental crust underlies the continents and their margins, and also small shallow regions in the oceans. In total it covers about 45 percent of the earth's surface and makes up about .3 percent of the mass. It is distinguished from oceanic crust and from the mantle under it by its physical properties and chemical composition. The horizontal boundaries between continental crust and oceanic crust are poorly defined; they are under not only the ocean's water column but also, in most places, a thick sequence of sedimentary rock. Seismic, magnetic and gravitational data indicate that the boundary is less than 10 kilometers across in some places but is several tens of kilometers across in others. Studies of the rock of oceanic and continental crust and the correlation of rock compositions and seismic velocities indicate, however, that the oceanic crust is characterized by the igneous rock basalt whereas the continental crust is an assemblage of igneous, metamorphic and sedimentary rocks enriched in elements such as potassium, uranium, thorium and silicon.

The vertical boundary between the mantle and the crust (both oceanic and continental) is called the Mohorovičić discontinuity, more commonly the Moho. It is a zone, less than one kilometer thick in some places but several kilometers thick in others, where the velocity of compressional seismic waves increases from about 6.8 kilometers per second in the crust to 8.1 in the mantle. The change in seismic velocity is caused largely by a change in the composition of the medium. Rocks of the mantle differ from rocks of the crust: they are poorer in silicon but richer in iron and magnesium.

Seismic studies of the Mohorovičić discontinuity indicate that the oceanic crust is typically from five to eight kilometers thick, whereas the continental crust varies from about 10 kilometers to more than 70. To a first approximation the crust behaves as if it were floating on the mantle. Oceanic crust is relatively thin and dense (3.0 to 3.1 grams per cubic centimeter); thus the parts of the earth's surface underlain by oceanic crust are usually far below sea level, at a depth ranging from 2,500 to 6,500 meters. Continental crust is thicker and is notably less dense (2.7 to 2.8 grams per cubic centimeter); thus the parts of the surface formed by continental crust lie near sea level or above it. The thickest parts of the continental crust usually underlie places of great elevation, such as the Himalayan and Andean mountain chains. Conversely, the thinnest parts of the continental crust usually lie below sea level at places such as the continental margins of the Atlantic.

Some important exceptions to this pattern are found at midocean ridges and some areas on the continents where volcanism is active and the crust is being stretched. In such places hot material from the deeper mantle rises to shallower levels, making the upper mantle hotter and so less dense than normal. The resulting buoyancy raises the surface elevation. The Basin and Range province of the western U.S. is a good example. The crust there is thin, but the surface elevation is nonetheless high.

Other exceptions to the pattern are found in areas with great topographic relief, where the crust bends downward over short horizontal distances, usually about 200 kilometers. Evidently the crust and the uppermost mantle deform like an elastic sheet to support the topographic load. One result is that long, linear troughs filled with sediment as much as eight kilometers deep form next to many great mountain chains. The troughs are underlain by crust of normal thickness. The crust has simply bent downward to support the weight of the mountains.

In general the rocks that form the continental crust fall into two major groups: widespread, relatively undeformed accumulations of sedimentary or volcanic rocks on the one hand and long, deformed belts of sedimentary, igneous and metamorphic rocks on the other. The belts are called orogenic belts, after the Greek *oros*, mountain. The first

LABRADOR FOLD BELT, shown in a Landsat image made above north Quebec, exemplifies the evolution of the continental crust: the belt was once a chain of mountains raised by the collision of two continents 1.8 billion years ago. Since then the mountains have eroded, exposing the deeper, mostly metamorphic and igneous rocks that the collision had deformed into myriad folds. Moreover, subsequent plate collisions have reshaped the continents. Nevertheless, the belt resembles the ones being formed by tectonic activity along continental margins today.



group is not ubiquitous on the continents, but where it is present it always overlies the second. In places such as the central U.S. it forms a thin veneer no more than a few kilometers thick. Elsewhere, along continental margins and in linear, circular or irregular depressions in the continents, it forms sequences of rocks that can be more than 10 kilometers thick.

The second group makes up the bulk of the continental crust. Its great variety of rock assemblages gives it a heterogeneity contrasting significantly with the relative homogeneity of the oceanic crust. Each belt in the second group evolved over a span of time as long as several hundred million years, and the ages of adjacent belts may differ by hundreds of millions of years, or even a billion years; hence each belt represents a different segment of earth history. Often the younger belts are oblique to the older ones, so that the younger truncate the older. In other places belts are parallel.

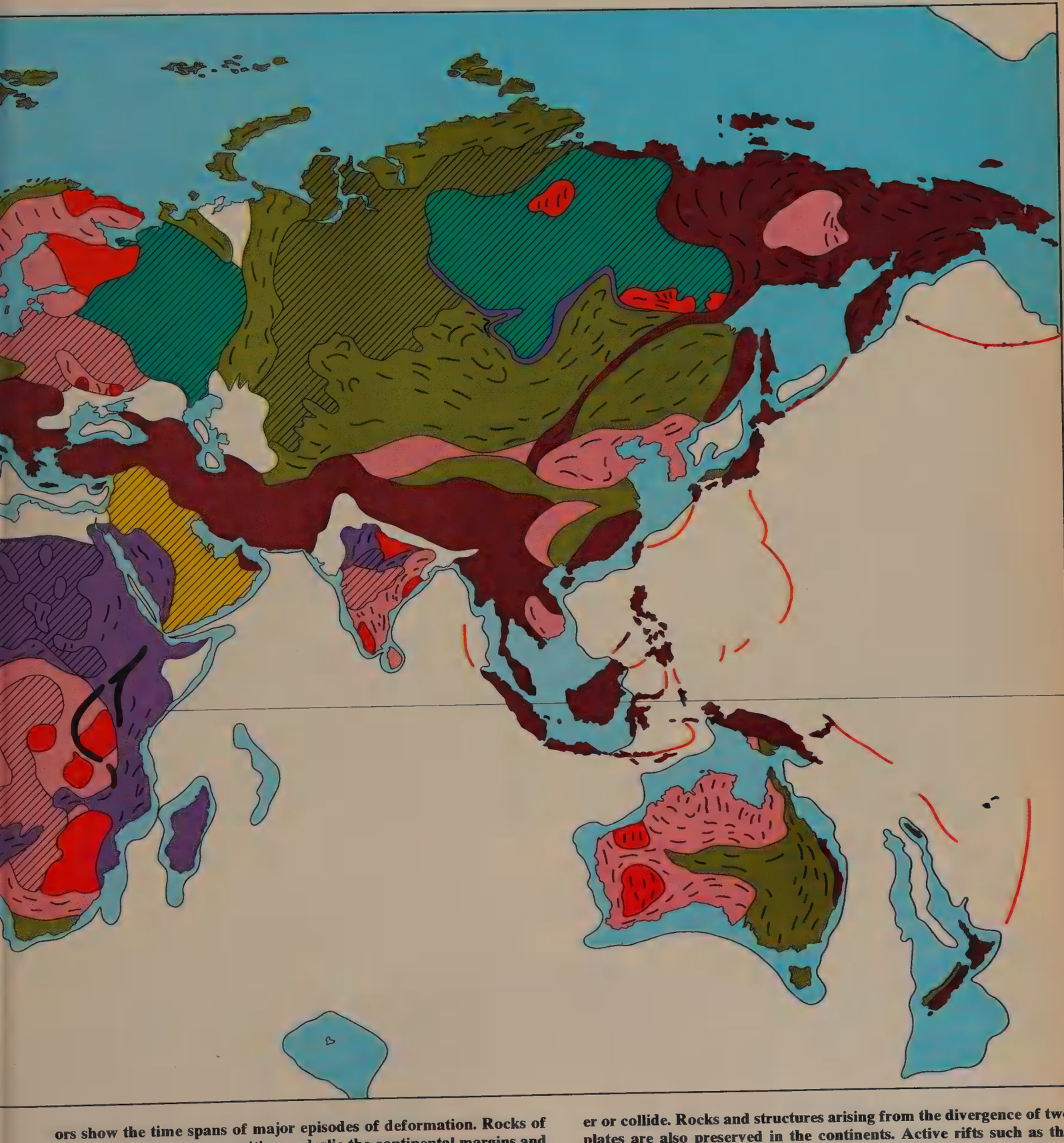
A detailed examination of the rocks in the belts shows that many older belts are similar to the ones formed in the more recent geologic past. They also



WORLDWIDE PATTERN OF BELTS of rocks deformed during different geologic periods accounts for the major proportion of the continental crust. In places the belts are overlain by sedimentary or volcanic rock or by ice (*black hatching*). Even then drilling and other methods of exploration confirm that the belts are present at depth. Each belt is complex: it records many individual episodes of deformation and usually includes rocks reworked from older belts. Col-

resemble the belts that the earth's tectonic activity is forming today. This offers the prospect that the study of modern rock formations and how they are currently deforming will yield understanding of the processes responsible for the formation of the older orogenic belts making up most of the continental crust. The theory of plate tectonics is crucial in such a venture, because the theory provides a framework in which rock assemblages and their deformation can be related to interactions of the plates that make up the entire crust. To be sure, the theory was developed largely from data gathered in the oceans, and its application to the study of continental crust has met with mixed results. Still, a modified version of the plate-tectonic concepts is a basis for understanding continental development.

The central concept of plate tectonics is straightforward: the outermost shell of the earth, the lithosphere, can be divided into six major plates and many smaller ones that move with respect to one another with velocities ranging from a few centimeters per year up to 20 or more. The plates consist of oceanic and continental crust together with



ors show the time spans of major episodes of deformation. Rocks of similar continental composition underlie the continental margins and parts of small oceanic plateaus and rises (*light blue*). In addition volcanic arcs (*orange*) built up on the oceanic crust form a type of crust that can be sutured to a continent when two plates slide past each other or collide. Rocks and structures arising from the divergence of two plates are also preserved in the continents. Active rifts such as the East African rifts, older rifts such as the Oslo rift in Scandinavia and ancient rifts such as the Athapuscow and Bathurst rifts in Canada, which are more than two billion years old, are examples (*gray*).

some of the underlying mantle; the Mohorovičić discontinuity lies within them. The plates are generally regarded as rigid bodies, so that most of their interactions are concentrated along plate boundaries, which can be zones of intense deformation. The boundaries can be classified into three basic types: divergent, transform and convergent. At divergent boundaries new oceanic crust is created; at transform boundaries it slides past the crust of a neighboring plate; at convergent boundaries it plunges into the mantle. The continental crust generally resists this subduction, largely because it "floats" on the mantle.

This basic scheme must be modified in several important ways if it is to help illuminate the evolution of continental crust. In the first place the continental crust participating in plate interactions often does not behave rigidly. In the upper lithosphere, therefore, the motions of plates may be absorbed partially or entirely by deformation within the continental crust. Much of the deformation can be attributed to forces arising at plate boundaries; hence the boundaries, which generally are narrow and well defined in oceanic crust, become broad and diffuse on the continents. The boundary between the Eurasian plate and the Indian plate, for example, is more than 2,000 kilometers wide in some of the places where the continental crust of one plate is in contact with the continental crust of the other.

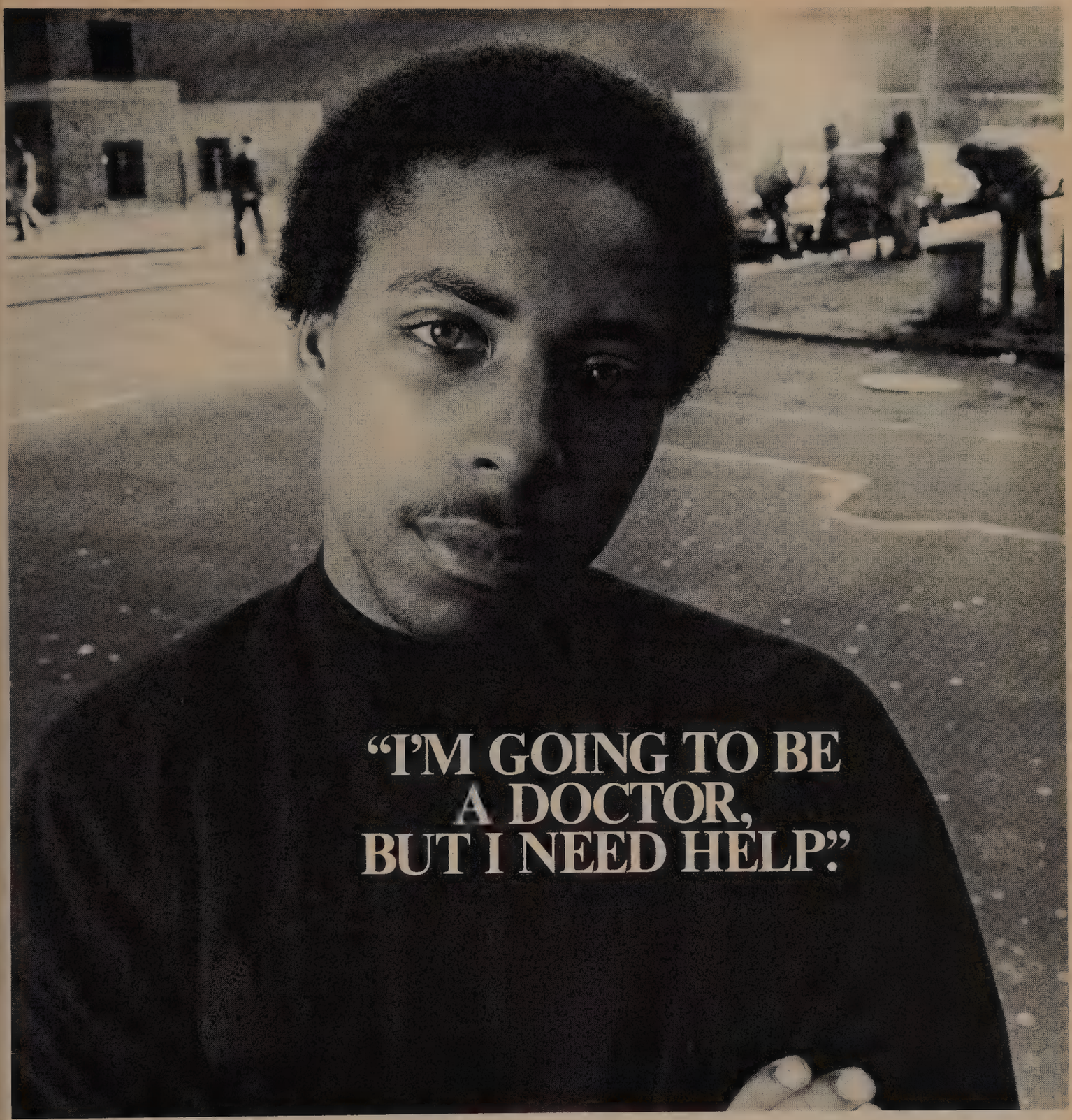
Second, the continental crust is markedly nonuniform in its mechanical properties, consisting as it does of belts of older rock and other preexisting structures that can localize new episodes of deformation. A zone of deformation extending well into a continent may thus form structures so greatly influenced by the anisotropies in the crust that the structures are hard to relate to the tectonic activity at the border of the plate. In such a zone of deformation it is difficult to define a plate boundary: the entire zone functions as the boundary. Commonly the older rock assemblages and structures of the continents have been subjected over time to the activity of many different plate-boundary systems. Therefore the record in the rock is often fragmentary and difficult to interpret. Under these circumstances the study of modern plate-boundary systems can show how modern and ancient plate boundaries evolve. Some examples from the three types of plate boundaries can serve as an introduction to the more complex patterns that result from the superposition of several tectonic episodes.

The divergence of two plates along a divergent plate boundary that crosses the continental lithosphere begins as the crust and its underlying lithospheric mantle become stretched and attenuated. Crustal faults develop in long, narrow zones, and within these zones the faulted crustal rocks subside differentially, forming great tilted blocks. Since the upper part of the mantle participates in the stretching, material from lower in the mantle (the hotter, more ductile level called the asthenosphere) rises to take its place, increasing the heat flow through the lithosphere. The result is the partial melting of the mantle and a characteristic volcanism of basaltic rock that is often alkalic (that is, rich in sodium and potassium).

Sometimes the divergence ends after only a few tens of kilometers of stretching, so that the zones of attenuated continental crust remain rifted scars in the



FLOTATION EQUILIBRIUM describes the relation of the continental crust to the underlying mantle. The crust, which is the lighter of the two, behaves as if it were floating; thus regions of great elevation, such as mountain chains, tend to be places where the crust is notably thick. Here two exceptions are shown. In the top drawing the crust near a mountain chain has flexed downward as if it were an elastic sheet supporting a weight. In the downward-flexed region, which has filled with sedimentary rock (*color*), the crust is thicker than one would judge from the elevation of the surface. In the bottom drawing, which shows the western U.S., a part of the mantle is hotter and so less dense than usual (*color*). Its buoyancy supports the crust, which is thinner than one would judge from the elevation of the surface.



A lot of qualified youngsters could not even consider a medical career if it weren't for one thing: Meharry.

Meharry, an independent 4-year medical college, educates far more black doctors and dentists than any other institution. In fact, without Meharry, there would be *half* as many of them practicing today.

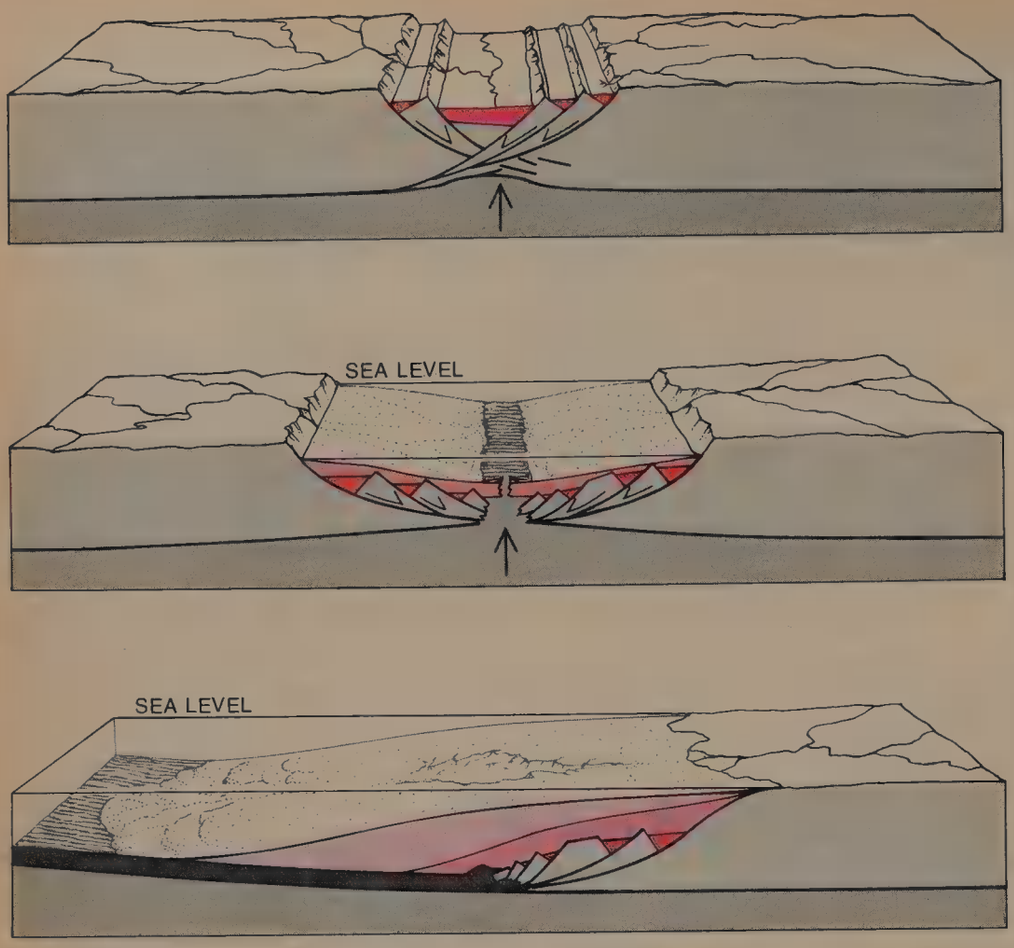
Of the 1,100 students, 49% come from families with annual incomes under $10,000. Tuition is $6,500 a year. Once over the pinch, 3 out of 4 Meharry graduates choose to practice in inner cities and rural areas.

To make more funds available for scholarships and student loans, and to maintain the quality of its faculty and facilities, Meharry is seeking to raise $27 million.

Won't you help? Determination alone won't make a doctor out of him. Only your dollars—and Meharry—will.

YOU HELP
WHEN YOU GIVE TO
MEHARRY
MEDICAL COLLEGE

Send your tax-deductible contribution, in any amount, to **Meharry Medical College, Second Century Fund, Nashville, TN 37208.**
The need is urgent, and your support is appreciated.



DIVERGENT PLATE BOUNDARY, where two plates move apart, causes extension and thinning of the continental crust. Initially (*top drawing*) the upper crust develops a series of brittle faults. Often the fault blocks rotate as they subside. The extension at deeper levels is less well understood. The subsidence (*middle drawing*) provides a site for the accumulation of nonmarine or shallow-marine sedimentary rocks (*medium colors*) and the rise of igneous rocks (*dark gray*). If the plates continue to diverge (*bottom drawing*), two continents result, and oceanic crust forms between them. The thinned margin of each continent subsides as it moves away from the zone of divergence; thus it is covered by unfaulted sedimentary rocks (*light color*).

continents. Some young, still active examples are the Rhine valley and associated rifts in central Europe, the East African rift valleys and the Rio Grande rift of the American Southwest. More ancient examples are the Oslo rift in southern Scandinavia, which is some 280 million years old, the Keweenawan rift of the central U.S., which is a billion years old, and the Athapuscow and Bathurst rifts in northwestern Canada, which are more than two billion years old.

In other instances the divergence continues. The attenuation of the crust thus becomes more extreme. It culminates in the separation of two bodies of continental crust and the formation of a new ocean basin, underlain by a widening span of oceanic crust. Each margin of the continental crust moves away from the region of divergence. The hot upper mantle moves with it. With time this hot upper mantle cools and contracts, causing the crust to subside. Meanwhile the faults that accompanied the stretching and attenuation of the crust become inactive. Sedimentary rocks begin to accumulate thick buildups above the thinned, subsiding crust. They also accumulate above the transition from continental crust to oceanic crust, forming the characteristic continental margin (called a passive continental margin) that flanks many ocean basins.

A profile through a rifted margin illustrates its evolutionary history. Thinned and faulted continental crust is overlain by a sequence of sedimentary and volcanic rocks deposited in faulted rifts during the initial phase of divergence. Those rocks are overlain in turn by a thick blanket of unfaulted sedimentary rock deposited during the latter, more gradual phase of subsidence. Well-studied examples of rifted margins are the Atlantic margins of the East Coast of the U.S. and the west coast of Africa. The margins generally have high temperature gradients during their early evolution, and so they are favorable sites for the maturation of organic matter into deposits of petroleum and natural gas. Occasionally the divergence of two continental bodies occurs near an older continental margin, and fragments of continent are rafted away to form small plateaus of continental crust partially or entirely submerged in the oceans and surrounded by oceanic crust. Examples include the Lord Howe Rise (with its highest part, New Zealand) in the southwestern Pacific and part of the Kerguelen and Mascarene plateaus in the Indian Ocean.

At transform boundaries, where two plates slide horizontally past each other along vertical or nearly vertical faults, crust is neither created nor destroyed. The horizontal displacement along the length of the boundary can measure hundreds of kilometers, even thousands. When the boundary crosses continental crust, the displacement is commonly distributed across a zone of faults as much as several hundred kilometers wide. Preexisting belts are shifted laterally, and parts of them can be rotated, greatly disrupting their original continuity. Moreover, offsets or bends in the faults can give rise to local regions of divergence or convergence.

Here are two examples. The Alpine fault in New Zealand is part of a transform system along the boundary between the Pacific plate and the Indian plate. It passes through a fragment of continental crust that was rafted away from Australia about 100 million years ago. The horizontal displacement along the fault is now about 400 kilometers, but the movement of the plates is not limited to displacement. In addition rock assemblages and structures created by activity at older plate-boundary systems have been rotated and bent, recording a total of about 1,200 kilometers of differential motion. It can be shown that the motion was purely transform about 40 million years ago and later became oblique, with components of both transform and compression. The compression has thickened the crust and raised a chain of high mountains: the New Zealand Alps.

The Dead Sea fault zone in the Middle East is a transform system connecting a divergent plate boundary in the Red Sea to a convergent plate boundary in the Taurus Mountains of southern Turkey. In places the fault zone steps to the west, cutting across the direction of transform motion and thereby creating small regions where the transform motion causes the crust to stretch, become attenuated and subside. The Dead Sea, the Sea of Galilee and the Gulf of Aqaba are all instances of such "pull-apart basins" along the fault. North of the northernmost basin the fault zone bends and steps to the east, producing the opposite result: the compression and thickening of the crust, which has raised the Palmyran Folds. In this way some of the northward motion of the Arabian plate with respect to Europe has been absorbed by convergence and shortening within the continental crust.

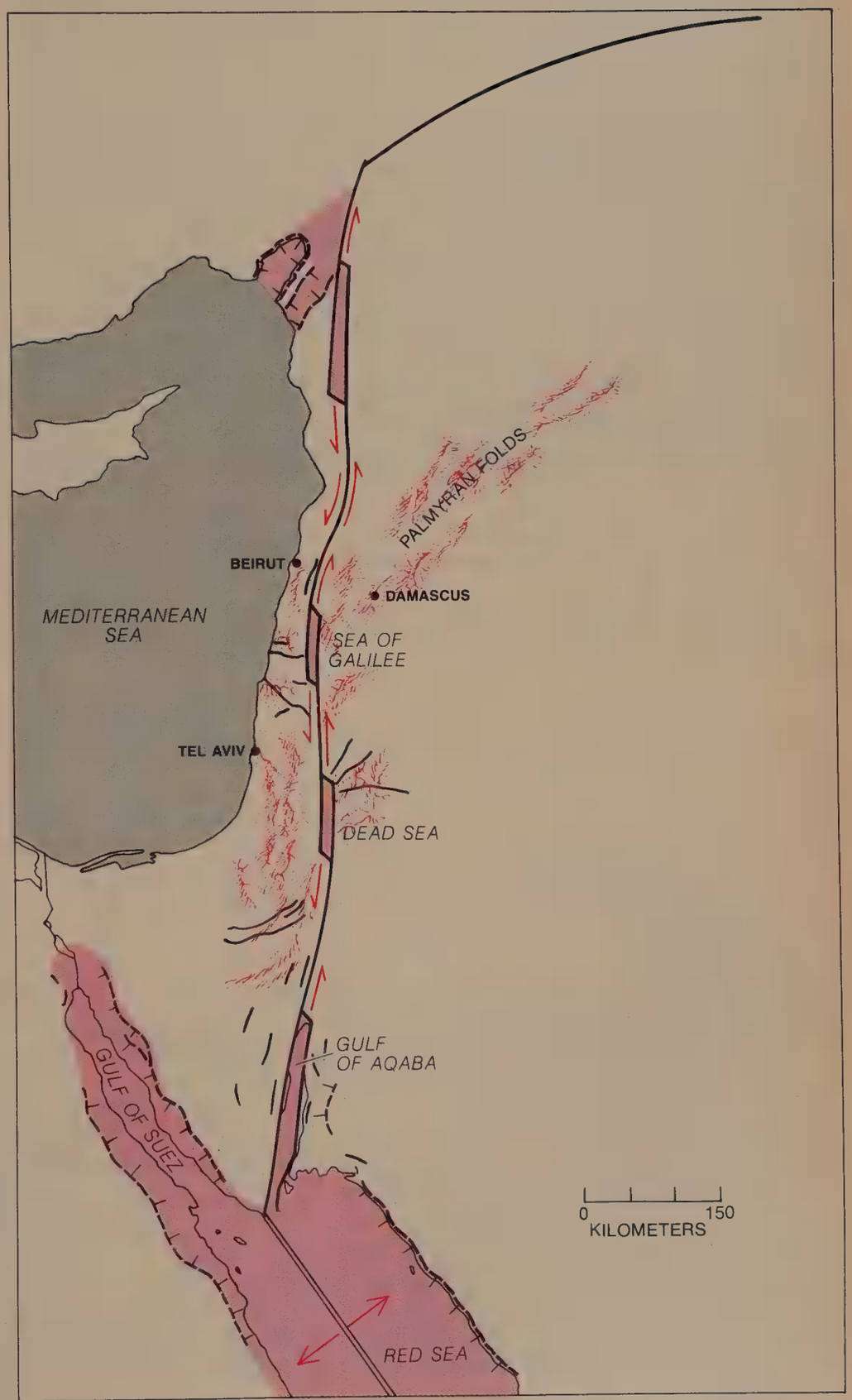
It is convergent plate-boundary sys-

tems that generate most of the continental crust. Among the three types of boundaries they are the most complex type, and in addition they deform the continental crust across the widest region. In the most usual configuration of a convergent boundary system a plate of oceanic lithosphere is subducted under an overriding plate of either oceanic or continental lithosphere. At increasing distance from the zone of subduction the overriding plate commonly shows a sequence of geologic features: first an accretionary wedge of folded and faulted sedimentary rocks and fragments of oceanic crust scraped from the top of the downgoing plate; then a topographic maximum (an "outer-arc high") formed by the most elevated parts of the accretionary wedge; then a fore-arc basin, which accumulates sediment from the adjacent elevations, and finally a volcanic arc, the most characteristic feature, fueled by magma rising from the subducted plate and the mantle just above it. If the overriding plate is oceanic lithosphere, the geologic features form what is called an island arc. If it is continental lithosphere, they form a continental-margin volcanic arc.

Behind the volcanic arc the overriding plate may be extended or compressed. Alternatively it may be relatively passive. If the plate is in extension, structures similar to the ones at divergent plate boundaries can form. A broad zone of stretching may thin the lithosphere and the crust, forming depressions such as the Aegean Sea of the eastern Mediterranean. If the extension of continental crust behind the arc proceeds until new oceanic crust is formed, a marginal sea will result. It will intervene between the greater part of the continental mass and the newly rifted fragment of continental crust, just as the marginal sea called the Sea of Japan intervenes between Asia and the islands of Japan. If the overriding plate is in compression, folds and faults arise in belts to accommodate the shortening and thickening of the crust behind the arc. In the Andes such features are found more than 800 kilometers behind the subduction zone.

At convergent zones new material from the mantle is added to the crust. In particular the subduction of oceanic lithosphere carries some ocean-floor sediments and the uppermost part of the oceanic crust downward into the mantle. The sediments and the crust contain water, and the water reduces the melting temperature of certain components of the subducted material. It also reduces the melting temperature of certain components of the mantle of the overlying plate. In short, the subduction of oceanic lithosphere causes partial melting at depth. The melted igneous material rises into the overlying rock. There it may cool and crystallize to form plutons: large subterranean igneous bodies. It may also reach the surface as lava or as explosive volcanic products such as pumice and ash. It is enriched in the elements common in continental crust; thus the partial melting advances the chemical differentiation of the outer part of the earth.

Geochemical studies of the igneous products show that they have had a complex, multistage history before coming to rest. In many instances the



TRANSFORM PLATE BOUNDARY, where two plates slide past each other, is exemplified by the Dead Sea fault zone in the Middle East. The crust at the east of the fault is moving north with respect to the crust at the west, and the relative displacement, which amounts to about 105 kilometers in the southern part of the zone, has opened a number of gulfs and seas, of which the Dead Sea is one. In addition parts of crust have shortened to form the Palmyran Folds.

igneous rocks have been contaminated by contact with older crustal rock, so that not all their volume represents new material derived from the mantle. Indeed, some igneous rocks are derived entirely from the melting of continental crust; hence they add nothing to the volume of the crust. The question of how much of the material added to the continents through igneous intrusion is new and how much is recycled is still unresolved.

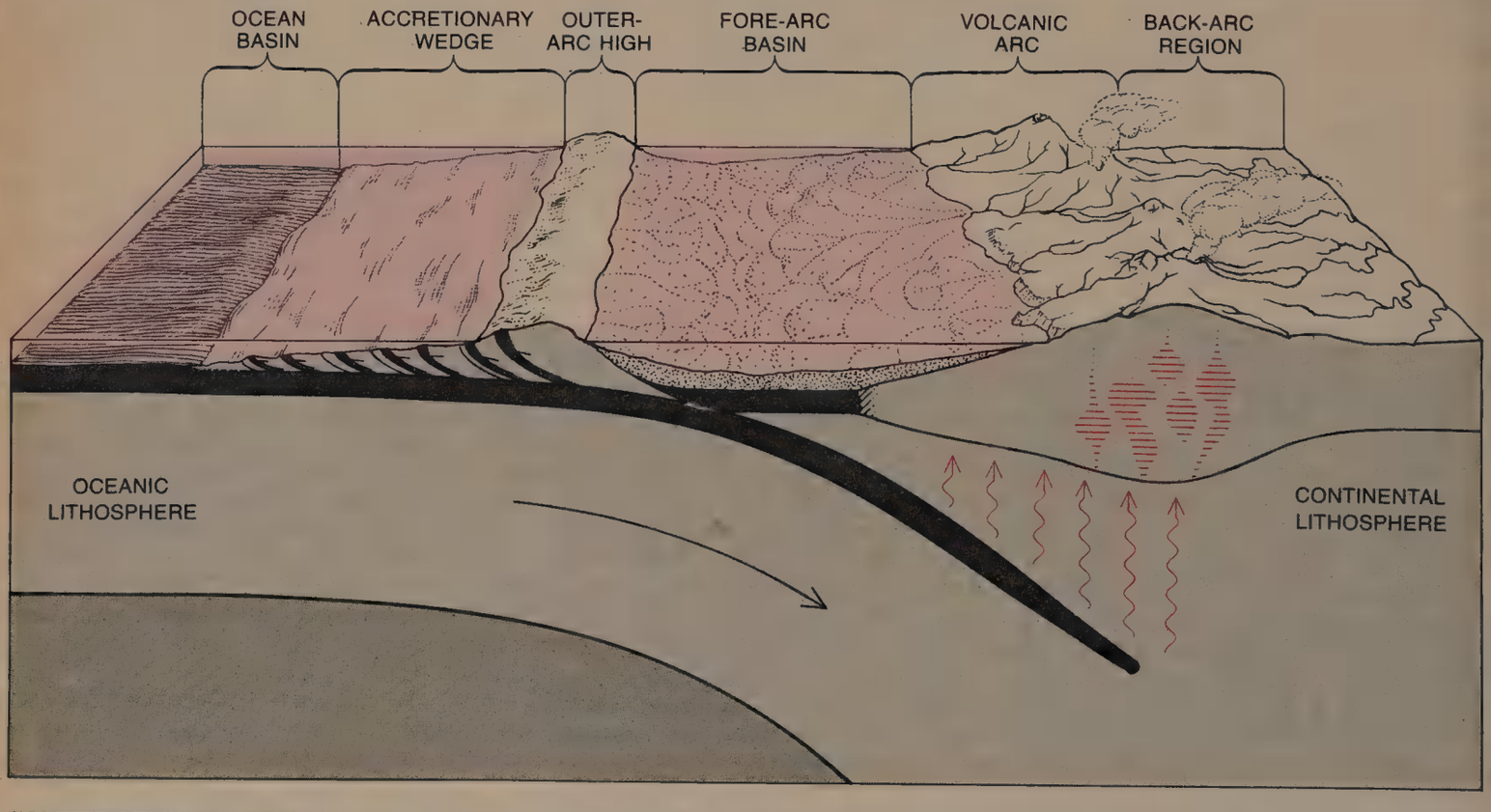
In any case the igneous intrusions increase the temperature within the lower crust, and the increase enhances the ability of the crustal rock to lose brittleness and deform in a ductile way. The structures in this ductile part of the crust may thus form large, complex folds. Moreover, the preexisting rocks may recrystallize into rocks with new mineral assemblages. The deformation and recrystallization may obscure and even obliterate the preexisting rock types and deformational patterns, so that it is difficult to elucidate the origin and evolution of the older rocks when erosion uncovers them at the surface of the earth. In general the regions of ductile deformation in the continental crust grade upward and laterally into regions of brittler deformation where the temperature remained lower.

Inevitably the convergence of plates leads to collisions between island arcs and continents. The arcs (along with oceanic islands and plateaus) are a transitional type of crust thicker and less dense than oceanic crust but not as thick and not as "light" as most continental crust. Nevertheless, the arcs, like continental crust, tend to resist subduction. Thus the elimination (by subduction) of the oceanic crust between an arc and a continent that are on converging plates leads to their becoming sutured together. The convergent motion during the collision may be perpendicular to the convergent boundary or it may be oblique to the boundary and have a component of transform motion. Where the motion is oblique the deformation within the collision system will have the characteristics of both convergent and transform boundaries.

The tectonic events at Papua New Guinea exemplify the geologic evolution caused by the collision of island arcs and continents. Here the convergence of the Australian plate and the Asian plate over the past 40 million years has driven one island arc or possibly two up over the edge of the Australian plate. The rock assemblages of the arc (or arcs) have been slivered and foreshortened, along with the assemblages of the ancient margin of Australia. Indeed, the northern edge of Australia has been shortened and thickened as far as 300 kilometers from the site of the collision. The disruption of the arc and of the continent has therefore been severe, but not so severe that the original relations among the rock assemblages cannot be deciphered.

Another type of geologic evolution is found where two continents collide along a convergent plate boundary. Such collisions are occurring today along the Alpine-Himalayan chain, where the Indian plate, the Arabian plate and the African plate are each colliding with the Eurasian plate. In the eastern Mediterranean the collision zone is more than 500 kilometers wide. There the geologic evidence suggests that several small fragments of continental lithosphere were swept together between the converging plates. The "buoyant" fragments stayed at the surface while oceanic tracts were subducted. The continued convergence in the region has deformed both the fragments and the margins of the plates, so that the collision system now extends across a broad zone.

One of the characteristics of collision systems, particularly those between continents, is subhorizontal decoupling, a process in which crustal sheets 10 to 20 kilometers thick slide over one another for tens or hundreds of kilometers. Such displacements stack and thicken different parts of the crust into a series of irregularly deformed and folded sheets, so that the rock assemblages and structures at depth in the crust cannot be pre-



CONVERGENT PLATE BOUNDARY, where two plates collide, is marked by a characteristic sequence of geologic features in the overriding plate. In the most usual configuration oceanic lithosphere (crust and upper mantle) is subducted under continental lithosphere. Fragments of oceanic crust and sedimentary rock scraped from the subducted plate form an accretionary wedge and an outer-arc high. Next comes a fore-arc basin, which accumulates sediments from the adjacent elevations, and then a volcanic arc, the most characteristic feature resulting from subduction. Some of the magma rising from the subduction zone solidifies in the crust. The back-arc region behind the volcanoes may show convergence (such as crustal faulting and folding) or divergence (such as crustal thinning and subsidence).

dicted from the assemblages and structures exposed at the surface. Another characteristic is that the convergence, which typically occurs along irregular boundaries in crust that is very anisotropic, causes complex motions of small crustal fragments within the convergent system. These local motions may be divergent, transform or convergent.

The most spectacular example of a convergent plate-boundary system active today is in Asia, where Peter Molnar of the Massachusetts Institute of Technology and Paul Tapponnier of the French National Scientific Research Council (C.N.R.S.) were the first to recognize that deformation extends across a region 3,000 kilometers wide. The collision some 50 million years ago between the Indian plate and the Asian plate represented a collision between the continental lithosphere of India and that of Asia. Since then a continued convergence that may total more than 2,000 kilometers has been absorbed principally by strain in the Asian plate. Broadly speaking, the Asian plate has absorbed the massive intracontinental deformation by compressional, transform and extensional faulting along young belts of deformation that generally follow older belts resulting from the activity of more ancient plate-boundary systems. To put it more simply, Asia has shortened longitudinally and extended latitudinally to accommodate the northward movement of India.

Meanwhile the northern edge of the Indian plate has broken into several gently dipping slabs whose pileup has thickened the crust and formed the Himalayas. The faulting in Asia extends nearly 3,000 kilometers from the collision boundary. Igneous activity in parts of the collision zone suggests that the deeper parts of the crust remain very hot today, creating an environment where rocks are recrystallizing and undergoing ductile deformation. Hence the rock assemblages and structures formed at earlier times are now being "overprinted."



BRITTLE DEFORMATION is characteristic of the rocks at shallow levels in a belt deformed by plate convergence. Hence sheets of rock tend to be thrust one over the other for tens or hundreds of kilometers. This photograph was made in the Spring Mountains of southern Nevada. The dark rocks are Cambrian limestones that are 550 to 500 million years old. They have been thrust from right to left over Jurassic sandstones lighter in color, which are 200 to 175 million years old. A well-defined thrust fault some 30 kilometers long marks their interface.

From studies of the youngest deformed belts, such as the Alpine-Himalayan belt, it is plain that plate boundaries evolve rapidly. Island arcs can be created, travel thousands of kilometers and collide with continents in only a few tens of millions of years. Small continental fragments can be rifted and collide with continents over similarly short spans of time. Thus the deformed belts that make up the underpinning of the continents represent a long, complex history of superposed plate-boundary systems.

The end of such a history probably comes when a region gets to be so remote from plate-boundary activity that it is no longer under its influence. In many instances this happens when an ocean basin is finally closed by a continent-continent collision. Parts of the Ural Mountains of the central U.S.S.R. are an example. First, about a billion years ago, two continental masses were rifted apart and a large ocean basin was created between them. The subsequent closing of that ocean generated island arcs, which collided with the continents at various times until the sequence was ended by the collision of two continents 250 million years ago. Much of the deformed belt resulting from that collision (namely the chain of the Urals) lies far enough from any younger plate-boundary activity for it to have suffered no further deformation.

This is not to say the entire belt is immune. Rifting in the south has created oceans younger than 250 million years, some of which have closed and created new mountain belts; hence the southern extension of the Urals has been overprinted by the Alpine-Himalayan belt. Moreover, the northern extension of the Urals reaches the Arctic Ocean, where future plate-boundary activity may be in store. Similar histories can be read in the geology of essentially all the continents in the crosscutting of older deformed belts by younger ones.

Of course, some histories are easier to read than others. In the youngest deformed belts the timing of tectonic events can be distinguished with an accuracy of better than one million years. In the older belts the timing becomes poorer with increasing age. The relative timing of many events can be determined, but the contemporaneity of events over large areas is difficult to establish. Thus it becomes a challenge to make accurate reconstructions of ancient plate-boundary systems. In addition the older belts are likelier to be ones where the once continuous locus of deformation has been disrupted. For example, the continuations of the ancient deformed belts in Australia are now found in India, Africa, Antarctica and South America.

The erosion of old deformed belts in regions where convergence has thickened the crust offers a valuable opportunity to study rocks that were once at deep crustal levels. It is remarkable just how deep erosion can reach. Erosion cuts into the terrain (a process taking tens of millions of years, even several tens of millions), progressively uncovering deeper levels, and since the crust is "buoyant," the removal of material from the top of it causes the remaining crust to rise. In effect it is timber pushed toward a saw. In this way rocks from depths as great as 30 or 40 kilometers come to be exposed at the surface.

By the time the rocks are exposed the

plate-boundary forces that shaped them have been long inactive. Still, the examination of the rocks enables geologists to infer the processes, the temperatures and the pressures that existed as the rocks deformed, and from such results a picture of the three-dimensional response of the continental crust to plate-boundary activity can be constructed. Only the deepest crustal levels escape exposure by erosion. Studies of inclusions in igneous rocks, chemical studies of igneous rocks and geophysical studies suggest that much of the deepest crust has a composition not unlike the composition of the shallower crust, except that the deep rocks are recrystallized so that their mineral assemblages are ones that are stable at high temperatures and pressures.

Some of the older deformed belts in the continents, including most of the Archean belts (the belts from 2.5 to 3.8 billion years old), have been difficult to interpret as plate-boundary systems analogous to modern ones. To be sure, the types of rocks in the older belts are similar to the types found in modern convergent systems. Their arrangements and structures, however, are somewhat different. Typically the older belts consist of bodies of volcanic and sedimentary rock, irregular or elongated in shape, along with large expanses of intrusive granitic and deformed metamorphic rocks that include mineral assemblages formed at high temperatures and pressures. The volcanic and sedimentary bodies resemble those found in modern island-arc settings. The only notable differences are that basaltic rocks are more abundant in the older belts and that some of the basalts are richer in magnesium and poorer in silicon than most modern basalts. On the other hand, many features of the younger belts appear to be missing in older ones. Thick, widespread sequences of shallow-water sedimentary rocks of the type that develop on modern passive continental margins are one example. Widespread vertical stackings of sheets of crustal rock that have been thrust one over the other are a second example.

Investigators of the Archean belts have taken two points of view: that the Archean belts are the result of plate motions whose geometry and intensity differed from those of modern plates, or that plate tectonics did not operate in the Archean, so that a mechanism of geologic evolution not observed today must be invoked instead. Although each view has its adherents, a modification of plate-tectonic theory can also be proposed. Perhaps the volcanic and sedimentary bodies in the Archean belts represent island arcs and their associated fore-arc basins and marginal seas, all of which were swept against small colliding continental nuclei. Larger, continental masses developed progressively, so that by about 2.5 billion years ago some orogenic belts began to take on a more modern appearance.

The rate at which continental crust has been formed remains a major question. On the one hand, the examination of orogenic belts indicates that many of them include a large amount of material derived from older belts by reworking or remelting. In addition many belts include rocks derived directly from the mantle by volcanic activity. Further still, many belts incorporate oceanic lithosphere. On the other hand, evidence suggests that continental material may sometimes be lost by being taken into the mantle. The studies show that orogenic belts vary widely in the balance of these processes. For example, the orogenic belts that were formed in north-central Canada between 2.5 and 1.8 billion years ago are proving to include much crust older than 2.5 billion years. In contrast, an orogenic belt of the same age in the southwestern U.S. contains little if any older crust. The studies do seem to indicate that the volume of continental crust has increased with time.

One of the very oldest Archean belts, about 3.8 billion years old, is in the continental crust of Greenland. Its rocks are sedimentary and igneous, and material in the sedimentary rocks has been derived in part from some older continental rocks. No direct evidence from that earlier era has been discovered, however, and so there remains a gap of 800 million years beginning 4.6

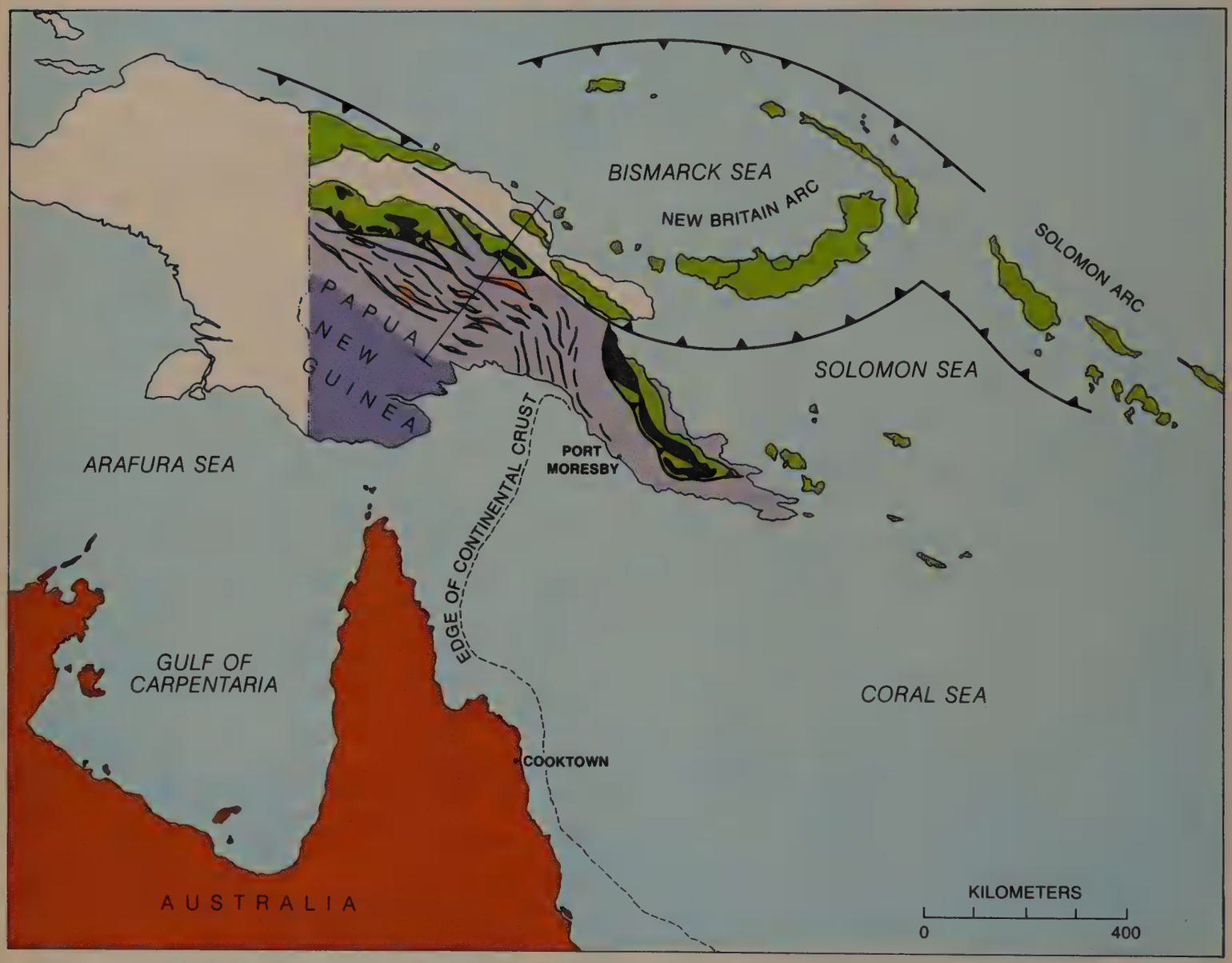


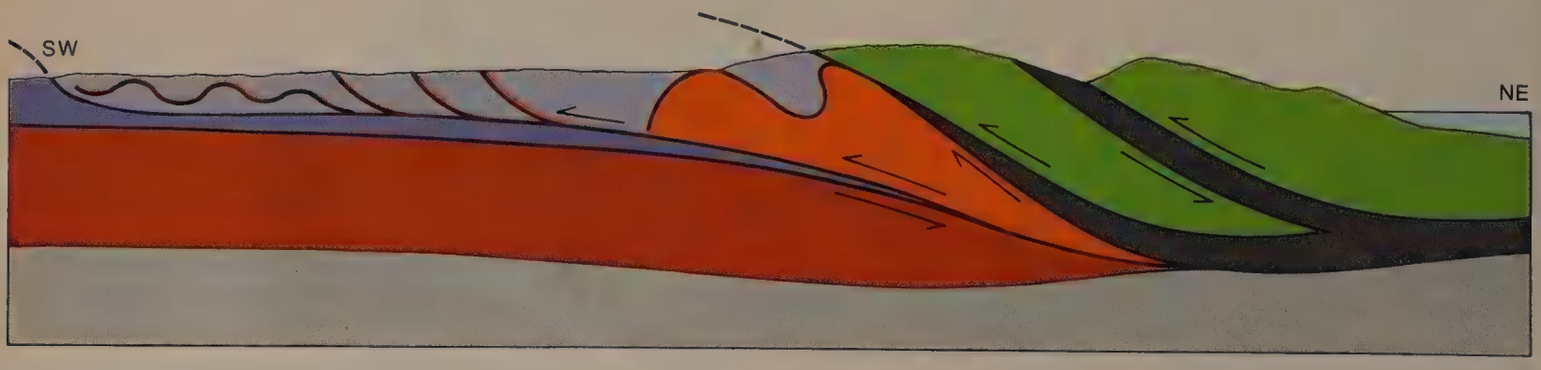
COLLISION OF CONTINENTS raised the mountain chains of eastern Europe and the Balkans. Fundamentally the African plate and the European plate converged, and the convergence shortened and thickened Europe by faulting its rocks into sheets and stacking them one above the other. The map shows the varying ages of the faults. Small barbs on each fault line face the overriding sheet. In general the stacking has a bilateral symmetry: the overriding sheets were driven northward in the northeastern half of the map and southward in the southwestern half. In the Pannonian Basin the crust has extended, thinned and foundered, so that the stacking is covered by sedimentary rock (*gray*). Broken lines indicate outcrops of deeper rock. The cross section suggests the complexity of the stacking; the colors employed in the section simply differentiate sheets of rock. The sheets are highly deformed; thus no individual sheet can be followed the width of the deformed belt resulting from the convergence.


WEST CARPATHIANS
RUSSIAN PLATFORM
EAST CARPATHIANS
TRANSYL-
VANIAN
BASIN
PANNONIAN BASIN
SOUTH
CARPATHIANS
DOBROGEA
MOESIAN
PLATFORM
DINARIC ALPS
SERBO-MACEDONIAN MASSIF
BLACK SEA
BALKAN
MOUNTAINS
RHODOPIAN
MASSIF
HELLENIC
ALPS
AEGEAN SEA
NE

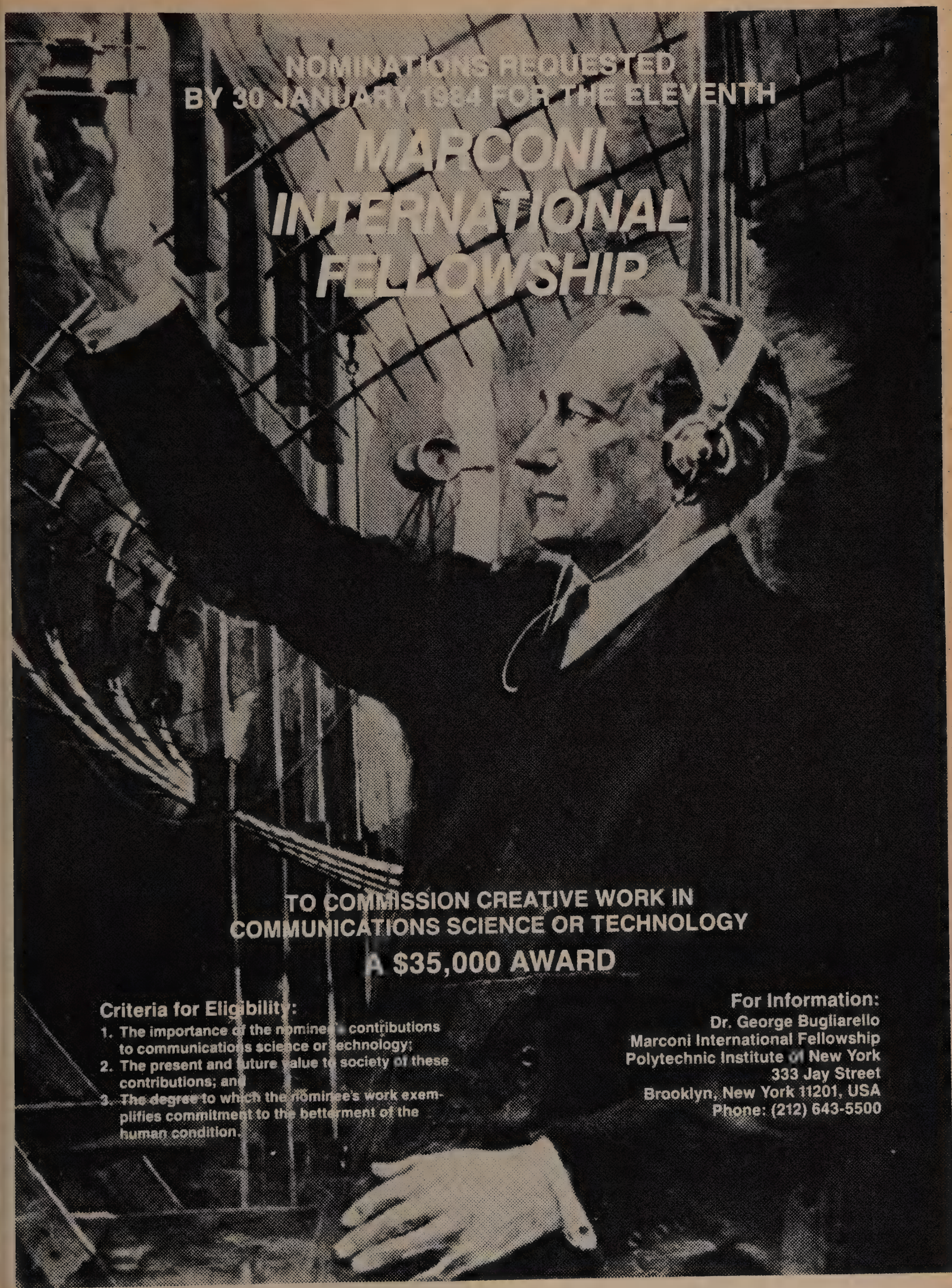
billion years ago, when the solar system is thought to have formed. In the later part of this gap the moon was intensely cratered; surely the earth was subjected to a similar bombardment. No signs of it remain. They were probably erased by the dynamic processes that have continuously created and reworked the continental crust.

Clearly the earth is an evolving body whose distribution of heat controls the motions, thickness and ductility of the lithosphere and the generation of igneous and metamorphic rocks. The generation of heat by radioactivity in the earth was probably about three times greater in the Archean than it is today. Temperature gradients in the earth were probably greater too, and that may help to explain at least some of the differences between Archean belts and younger ones. Until the variables that affect the formation of orogenic belts are better understood, the Archean belts will remain a major challenge to the understanding of how the continents came into existence.





COLLISION OF ISLAND ARCS with the northern part of the Australian continental crust over the past 40 million years has sutured some arcs (*green*) and also some remnants of oceanic crust (*gray*) onto the northern part of Papua New Guinea. The arcs were probably formed by the subduction of the Australian plate under the Pacific plate; then the convergence of the plates carried Australia into the zone of the subduction. The continued convergence has now folded and faulted sedimentary rocks along the margin of Australia (*purple*) and also the ancient crust of Australia itself (*orange*). Darker shades signify rock that escaped deformation. The cross section is schematic.


NOMINATIONS REQUESTED
BY 30 JANUARY 1984 FOR THE ELEVENTH
MARCONI
INTERNATIONAL
FELLOWSHIP
TO COMMISSION CREATIVE WORK IN
COMMUNICATIONS SCIENCE OR TECHNOLOGY
A $35,000 AWARD
Criteria for Eligibility:
1. The importance of the nominee's contributions to communications science or technology;
2. The present and future value to society of these contributions; and
3. The degree to which the nominee's work exemplifies commitment to the betterment of the human condition.
For Information:
Dr. George Bugliarello
Marconi International Fellowship
Polytechnic Institute of New York
333 Jay Street
Brooklyn, New York 11201, USA
Phone: (212) 643-5500

HOW TO REACH THE PEOPLE WHO MAKE THE FUTURE HAPPEN IN GERMANY



SPEKTRUM DER WISSENSCHAFT is the German edition of SCIENTIFIC AMERICAN which attracts a young audience of affluent professionals who have come to the top because of their technical expertise. Almost half of them hold top management job titles. The fortunes of major corporations in Germany increasingly depend on technically sophisticated people. SPEKTRUM DER WISSENSCHAFT keeps these people in touch with the advances in science that drive the growth of Germany's industry.

SCIENTIFIC AMERICAN speaks the languages of more than half the world's population. We are the one publication in the world today providing efficient coverage of technology-based management. We reach men and women in industry, whose qualification to make technical decisions places them in key positions in their country's government and industry.

SCIENTIFIC AMERICAN in eight languages: English, Spanish, Italian, French, German, Japanese, Chinese and now Russian has gathered in its audience three million people who make the future happen around the world.

For more information on our German-language edition contact:

Claus G. Firchow
SPEKTRUM DER WISSENSCHAFT
Moenchhofstrasse 15
6900 Heidelberg
FEDERAL REPUBLIC OF GERMANY
Telephone 011-49-6221-46037

or in New York
John Kirby
V.P./Advertising Director
SCIENTIFIC AMERICAN
415 Madison Avenue
New York, New York 10017
Telephone 212-754-0262

THE INPUT/OUTPUT STRUCTURE OF THE UNITED STATES ECONOMY



WHAT MAKES THE U.S. ECONOMY TICK?

The editors of SCIENTIFIC AMERICAN have prepared a wall chart displaying for the 1980's the Input/Output Structure of the U.S. Economy based on the latest interindustry study from the U.S. Department of Commerce.

The SCIENTIFIC AMERICAN Input/Output wall chart does for economics what the table of elements does for chemistry. It answers at a glance questions about the linkage between the microeconomics of the firm and the macroeconomics of the system; about the web of technological interdependencies that tie industry to industry; about the industry-by-industry direct and indirect consequences of swings in public and private spending; about the impact of change in technology, and about any other topic you can think of. You are rewarded by surprise as well as by confirmation of your hunches. For teaching and practical and theoretical studies, here is a powerful, graphic tool.

In the familiar format of the SCIENTIFIC AMERICAN Input/Output wall charts for the 1960's and 1970's, the wall chart for the 1980's measures 65″ × 52″ and is printed in eight colors. Each of the nearly 10,000 cells in the 97-sector interindustry matrix shows (1) the interindustry commodity flow, (2) the direct input/output coefficient and (3) the "inverse" coefficient. Where the direct input/output coefficient exceeds .01, the cell is tinted in the color code of the industrial bloc from which the input comes. This device, combined with triangulation of the matrix, brings the structure of interindustry transactions into graphic visibility.

A supplementary table displays, industry by industry, the capital stock employed; the employment of managerial, technical-professional, white-collar and blue-collar personnel; the energy consumption by major categories of fuel, and environmental stress measured by tons of pollutants.

The editors of SCIENTIFIC AMERICAN are happy to acknowledge the collaboration, in the preparation of this wall chart, of Wassily Leontief, originator of input/output analysis—for which contribution to the intellectual apparatus of economics he received the 1973 Nobel prize—and director of the Institute for Economic Analysis at New York University.

Packaged with the chart is an index showing the BEA and SIC code industries aggregated in each of the 97 sectors.

SCIENTIFIC AMERICAN
415 Madison Avenue, Dept. I/O, New York, N.Y. 10017

YES, send me the new SCIENTIFIC AMERICAN 1980's INPUT/OUTPUT wall chart—printed in eight colors on heavy paper measuring 65″ × 52″.

I understand my chart(s) will be shipped prepaid, uncreased*, flat-rolled and in a sturdy mailing tube. (Allow 4-6 weeks for delivery)

I enclose a check for $____ for ____ chart(s) at $35 each, saving me $5 each in handling charges.

Bill me for $____ for ____ chart(s) at $40 each.

Name_____

Company_____

Street Address_____

City, State, Zip Code_____

*Charts shipped international require one fold. 9/83

Make check payable to SCIENTIFIC AMERICAN. Please add sales tax for California, Illinois, Massachusetts, Michigan and New York.

The Ocean

The chemistry of the ocean, whose constituents interact with those of the air and the land to support life and to influence the climate, is now known to have undergone changes since the last glacial epoch

by Wallace S. Broecker

Water and air interact with the lithosphere to provide chemical support for the community of living things. The ocean and the atmosphere constitute a single system that functions as an integrated chemical plant. In attempting to reconstruct the history of the earth, geologists have tended to think of the chemical composition of the ocean and the atmosphere as being essentially static through time. The reason is that geochemists have been slow to come up with reliable paleochemical indicators.

In the past 10 years there have been important breakthroughs in the study of ocean chemistry. There is now reasonably firm evidence, at least in the case of those constituents of the seas' salts that are heavily incorporated into organisms and processed by them, for significant changes in concentration over time. Some of these fluctuations (the best-documented ones) seem to have been in tune with major cycles of glaciation that have dominated the earth's environment for the past million years or so. Others were surely related to changes in the architecture of continents and ocean basins caused by movement of the great lithospheric plates. Still others may be effects of such geologic catastrophes as the great asteroid or comet impact that, it now appears, marked the end of Cretaceous time.

Fluctuations in the ocean's chemistry are worth studying for what they tell about variations in the planet's environment in the past. Ocean chemistry has affected the environment primarily by driving changes in the atmospheric content of one constituent in particular: carbon dioxide (CO_2). Much of the solar energy impinging on the earth is reradiated from the surface in the form of infrared radiation, which is strongly absorbed by carbon dioxide and so is trapped in the lower atmosphere as heat. The partial pressure of carbon dioxide in the atmosphere therefore influences temperature and hence also rainfall, winds and the extent of ice cover. It is well known that man's activities, notably the clearing of forests and the burning of fossil fuels, are increasing the atmosphere's carbon dioxide content, but only recently has it become clear that there were carbon dioxide changes in the past as well.

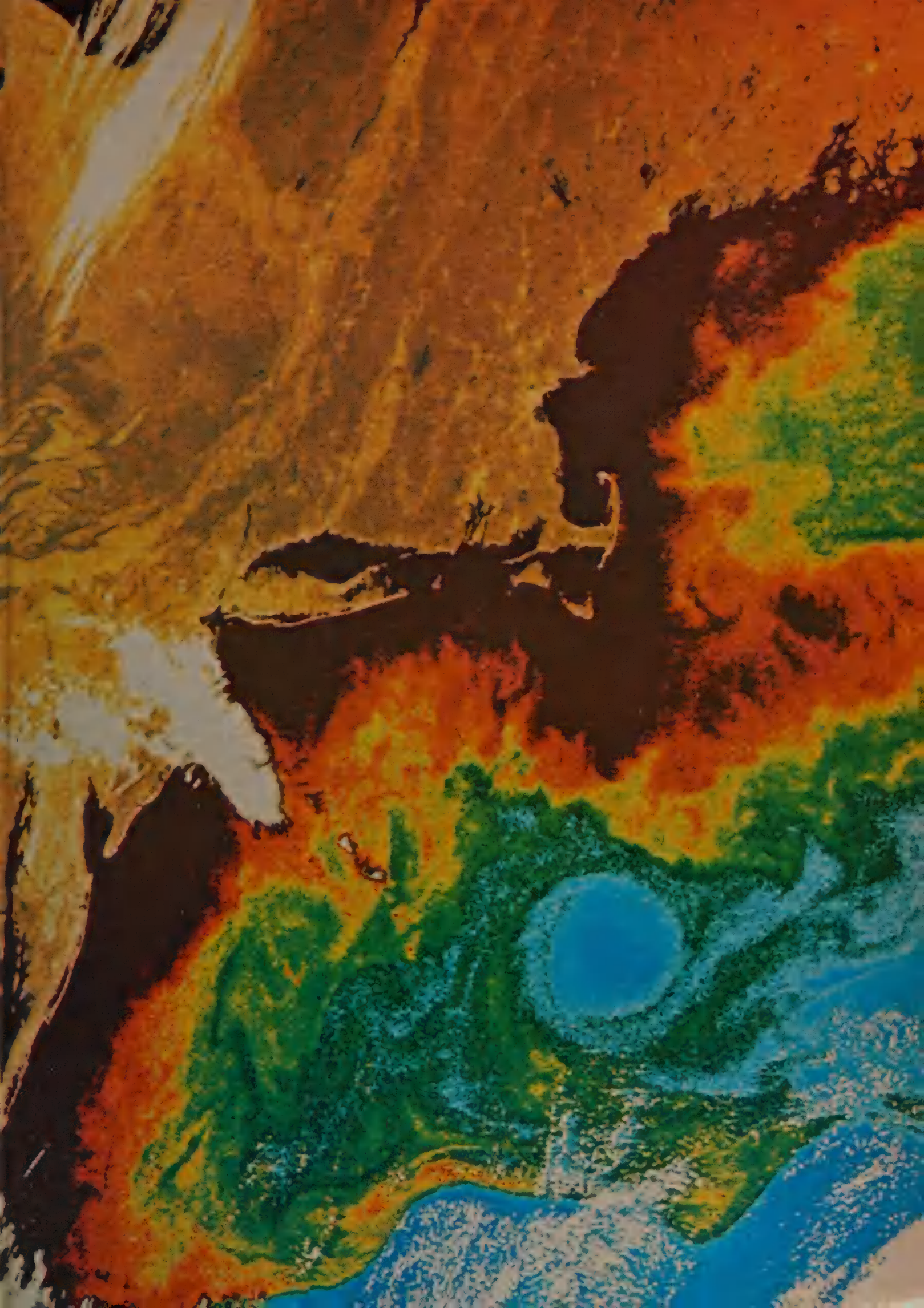
The major immediate source for the carbon dioxide in the air is the ocean, which at present holds, dissolved in one form or another, 60 times as much carbon as there is in the atmosphere's carbon dioxide. The transactions controlling the partitioning of carbon between the sea and the air depend on the detailed chemistry of those constituents of sea salt that are heavily utilized by marine organisms. To know the ocean's role in the carbon cycle it is necessary to study mechanisms controlling the concentrations of such constituents in the ocean: the interconnected feedback loops whereby those concentrations are driven toward dynamic equilibrium. Somehow the ocean must cope with the heavy demand of marine organisms for these critical ingredients, maintaining a balance between input from either erosion or volcanic activity and ultimate loss by burial in sediments.

The control mechanisms can best be understood if one thinks of the ocean as a giant two-reservoir aquarium. The upper reservoir interacts with the air and the land. Water evaporates from the sea surface, condenses and falls onto the land as rain. The percolation of rainwater through the land dissolves and carries into the ocean some of the seawater constituents I shall be discussing. Others are supplied by volcanic activity, both on the land and on the sea floor. The upper reservoir is lighted by the sun, whose energy is converted by plants (mainly microscopic phytoplankton) into organic matter; the plant matter supports a food chain of animals and bacteria. Debris from these organisms falls from the upper reservoir into the dark lower reservoir, where it nourishes other animals and bacteria. Some residues fall to the ocean floor, where they are decomposed by worms and bacteria. A small fraction of the residues remain unconsumed and are buried, and so the material dissolved from the land or injected into the atmosphere or the ocean as volcanic gas is ultimately deposited as sediment.

To keep things simple I shall deal here only with dissolved oxygen, calcium ion, phosphate and three carbon compounds: carbon dioxide, carbonate ion and bicarbonate ion. The choice of these constituents reflects the chemical requirements of marine organisms. Marine animals need dissolved oxygen for respiration. For their biochemical functions all organisms need carbon and phosphorus (along with hydrogen and oxygen, which are of course superabundant in water). Some organisms house themselves in chambers made of calcite (calcium carbonate).

Calcium enters the aquarium in the form of the doubly positively charged calcium ion (Ca^{++}) dissolved in river water and is lost to sediments as calcium carbonate ($CaCO_3$). Carbon enters through the erosion of carbonate miner-

CHLOROPHYLL CONCENTRATION in waters off the U.S. and Canadian coast is mapped in the false-color image on the opposite page. The nearly lifeless warm water of the Gulf Stream (and of a circular eddy pinched off from it) is bright blue, and successively increasing chlorophyll concentrations are represented in light blue, green, yellow, reddish brown and dark brown. Chlorophyll is manufactured by phytoplankton, the microscopic marine plants at the base of the oceanic food chain. As is described in this article, these microorganisms have a major role in the control mechanisms that regulate the chemistry of the ocean-atmosphere system. Data for the image were collected by the coastal-zone color scanner carried aboard the satellite *Nimbus 7*. It records energy radiated from surface waters in spectral bands that are more or less strongly absorbed by chlorophyll. The experimental team for the color scanner is headed by Warren A. Hovis, Jr., of the National Oceanic and Atmospheric Administration.

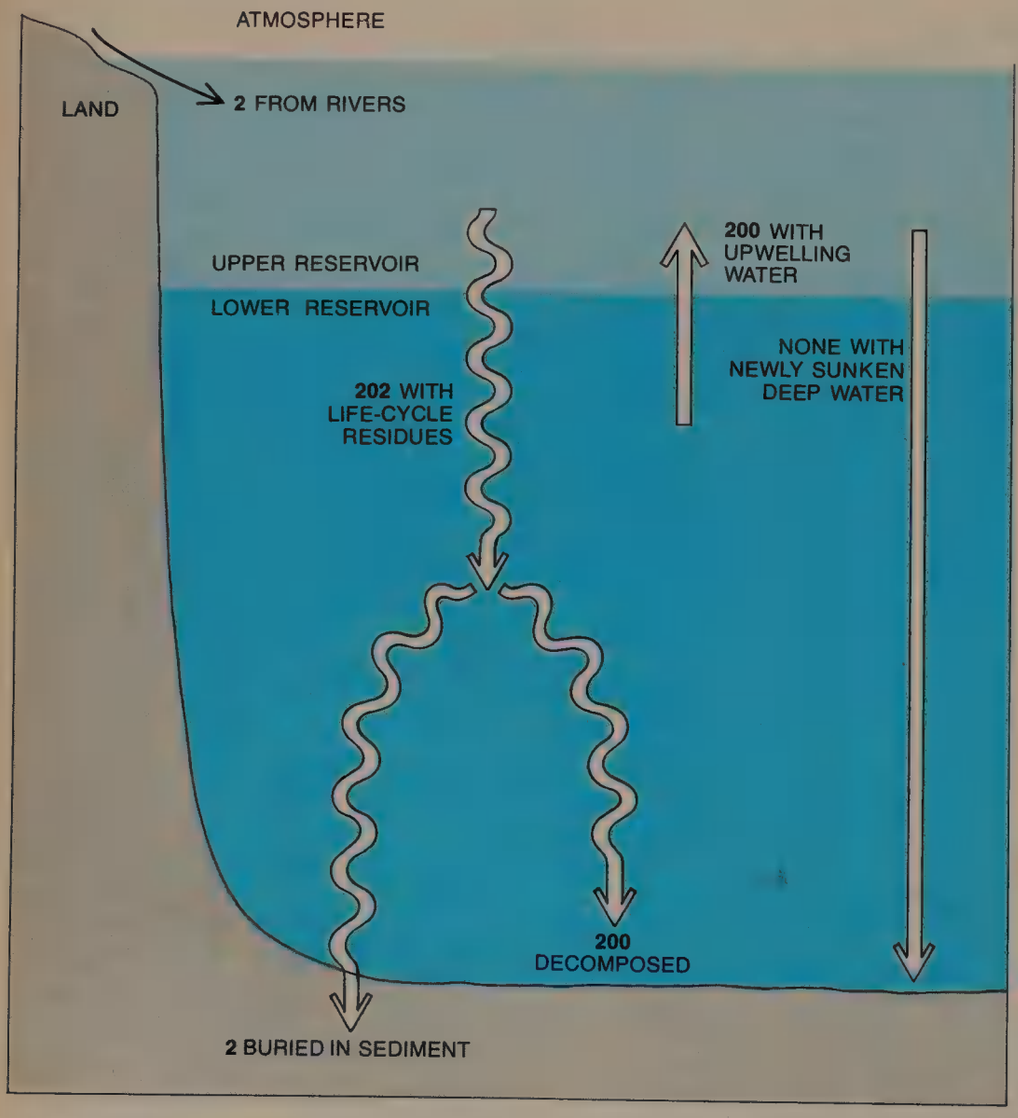


als (as the doubly negatively charged carbonate ion, CO_3^{--}) and from volcanic sources (as carbon dioxide gas); it is lost both as calcium carbonate and as organic matter in soft-tissue residues. Phosphorus is supplied as phosphate (PO_4) by the erosion of phosphate minerals and is lost to sediment in soft-tissue residues. Dissolved oxygen (O_2) is generated by plants and consumed by animals. In this simplified aquarium world four control loops are needed to keep the loss of ingredients in balance with their supply. A fifth control loop is needed to maintain electrical neutrality by balancing positive and negative ions.

I shall not deal with nitrogen, even though it is a key limiting ingredient for organisms. Two justifications for this simplifying omission are the superabundance of dissolved nitrogen gas and the observation that the ratio of nitrate (NO_3) to phosphate in seawater is everywhere about the same. Apparently organisms that are able to convert nitrogen gas into nitrate (the form most readily utilized by most plants) manage to maintain in the sea a nitrate-to-phosphate ratio that meets the biochemical demand. Since there is no such means by which living things can enhance their supply of phosphorus, it is phosphorus that becomes the ultimate limiting ingredient.

One must also consider the "residence time" of each of the ingredients in the ocean, which is to say the ratio of its total inventory in the ocean to its rate of supply (or of loss). In today's ocean the residence time, thus calculated, for phosphorus atoms is about 100,000 years, for carbon atoms about 165,000 years and for calcium atoms about a million years. The sizable difference between the residence time of phosphorus and carbon on the one hand and of calcium on the other allows for further simplification when one considers chemical changes associated with glaciation. Individual glacial episodes average some 50,000 years in duration. On such a short time scale the calcium content of sea salt cannot change significantly. Oxygen does not lend itself readily to a residence-time calculation, but there is so much of it in the atmosphere that its abundance is not likely to change significantly even in a million years. In considering chemical changes since the last glaciation, therefore, one needs to deal only with carbon and phosphorus (and charge balance).

Finally, the mechanisms of water transport between the two reservoirs of the ocean must be taken into account. The two reservoirs are separated by the main thermocline: the boundary region separating the cold waters (maintained by the sinking of dense water from the surface in polar regions) from the warm surface waters (maintained by sunshine). The strong difference in density between the two realms limits the rate at which they mix, so that the mean time spent by a water molecule in the cold-water compartment between trips to the warm surface ocean is about 1,000 years. A typical phosphorus atom therefore makes about 100 trips back and forth across the thermocline in the course of its 100,000 years in the ocean. A carbon atom makes 165 such trips in the course of its residence time, a calcium atom 1,000.



TWO-RESERVOIR OCEAN consists of a warm surface compartment and a much larger body of denser, colder water. The movement through the ocean of chemicals that are critical to marine life is represented here by an idealized phosphorus cycle. Phosphorus eroded from the land is delivered to the sea by rivers; eventually it is removed by being buried in sediment as part of undecomposed organic residues. Loss matches input (shown here in tens of billions of moles per year). Almost all the phosphorus reaching the warm surface ocean is taken up by marine plants and carried to the deep sea with falling residues, most of which decompose, releasing phosphorus into solution; the regenerated phosphorus returns to surface with upwelling water.

Although carbon and phosphorus move with upwelling water from the deep ocean to the surface layers, in their return trip to the deep sea these elements have a different option: they can fall, chemically bound within a fragment of organic matter. Almost all the phosphorus that moves up across the thermocline in upwelling water falls back to the cold-water compartment in organic particles, whereas only about an eighth of the upwelled carbon takes the particulate short cut; the remainder returns dissolved in sinking water. The reason for the discrepancy is that the ratio of carbon to phosphorus in sea salt is some eight times higher than the ratio in organic debris, so that in exhausting the available phosphorus the community of living organisms in the upper ocean uses up only 12.5 percent of the dissolved carbon. In removing carbon to form their tissues marine plants reduce the partial pressure of carbon dioxide in the surrounding water and thus also in the atmosphere. The effect is so large that if life in the ocean were suddenly suspended, the carbon dioxide pressure would rise significantly in surface water, and it would about triple in the atmosphere.

The history in the ocean of an average phosphorus atom can be traced from

when it issues from the river mouth and enters the surface sea. Soon it is taken up by a plant. After being traded several times among the plant, animal and bacterial populations of the surface water, it is trapped in a residue (such as a fecal pellet) that falls into the deep sea. Of 100 phosphorus atoms bound in falling residues 99 are released back into solution; one is removed from the sea to permanent storage in the sediment. The redissolved atoms eventually find their way back to the surface and repeat the cycle. The average phosphorus atom is therefore dissolved in the deep sea in 100 separate 1,000-year terms, which are separated by brief interludes of only a few months during which it serves the biochemical needs of organisms in the warm surface ocean.

A match between the removal of phosphorus (with undestroyed residues) and phosphorus input to the sea (in river water) is maintained by adjustments in the ocean's phosphorus content. The higher the phosphorus content is, the more atoms reach the surface with upwelling water, the more residues fall from surface water and the more residues are sequestered in accumulating sediment. If some perturbation were to cause the loss of phosphorus to exceed input, the phosphorus inventory of the ocean would begin to decrease; there would be corresponding decreases in the rates of production and burial of residues. The mismatch between output and input would gradually be reduced until it eventually disappeared. The time scale for the adjustment would be of the same magnitude as the oceanic residence time of phosphorus atoms, or about 100,000 years.

How might the system be perturbed? Suppose a climatic change were to double the rate of exchange of water between the surface and the deep ocean. The delivery rate of phosphorus to the surface would double, as would the production and burial rates of residues. If the situation persisted, the ocean's phosphate content would decrease until it reached half of the initial value and the balance between input and output would have been restored. If, on the other hand, the river-input value were halved, the ocean's phosphorus inventory would eventually be halved. And if there were a 1 percent reduction (from 99 percent to 98) in the efficiency with which residues falling into the deep sea are destroyed, releasing phosphorus to dissolution, the phosphorus inventory would also be halved.

In my idealized ocean phosphorus is the limiting ingredient for life, and once its concentration is fixed so too is the rate at which residues are generated to rain down from the surface sea. This must be roughly correct, but ecological factors could alter the ratio of carbon to phosphate in soft-tissue residues or the ratio of falling calcite to falling soft-tissue residues. Putting such factors aside, one can assume that once the rate at which phosphorus rains down from the surface is set, so too is the total rate at which carbon rains down (in both soft tissues and calcium carbonate). At present, however, only about a sixth of the carbon delivered to the ocean by erosion and volcanism is lost with soft-tissue residues. The remaining five-sixths must leave the sea as calcium carbonate. How is the flow of this carbon regulated?

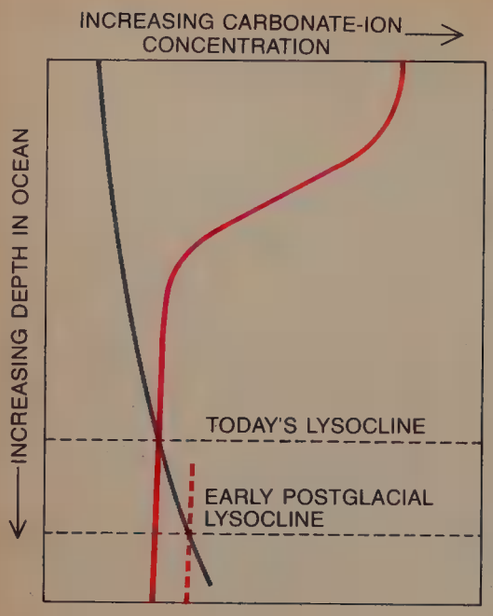
Whereas for phosphorus a balance between loss and gain is achieved by regulation of the rate of production of phosphorus-bearing residues, the balance for carbon is controlled by the dissolution of calcium carbonate residues. To understand this scheme of control by dissolution it is necessary to understand how the distribution of carbonate ion and its saturation level vary with depth. The carbonate content decreases with depth because carbon dioxide is removed from surface water as a result of photosynthesis (increasing the carbonate-ion concentration) and is returned to solution in deep water (decreasing the carbonate concentration). The saturation level for carbonate, on the other hand, increases with depth because calcite's solubility increases with pressure.

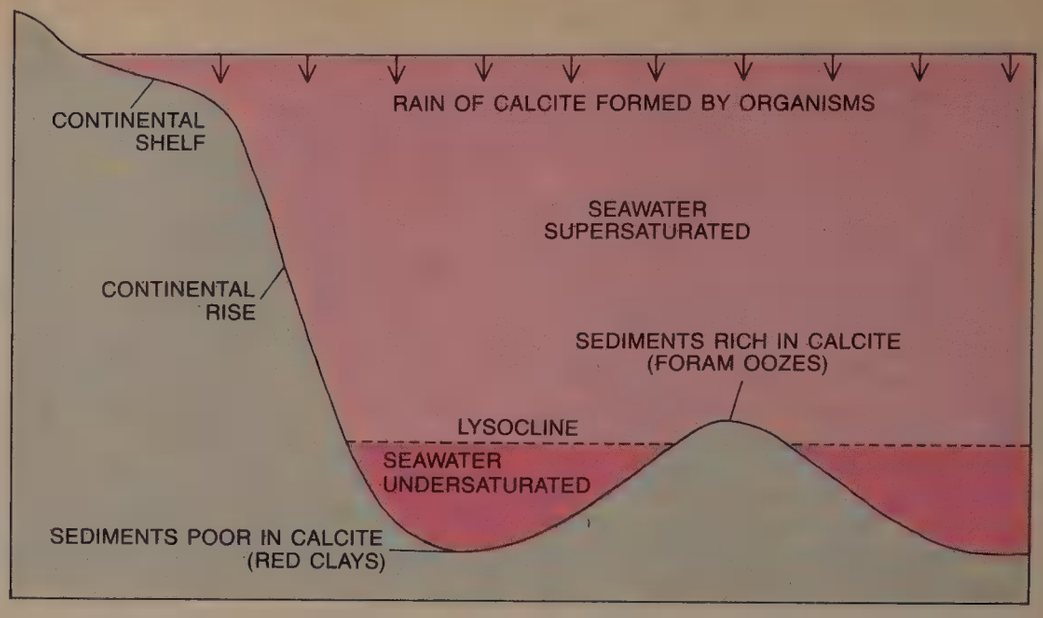
	INPUT TO OCEAN		SEA SALT	MARINE-ORGANISM DEBRIS	LOSS TO SEDIMENT	RESIDENCE TIME (THOUSANDS OF YEARS)
PHOSPHORUS	(EROSION)	1	1	1	(ORGANIC MATTER) 1	100
CARBON	(CARBON DIOXIDE, VOLCANISM)	100	1,000	125	(ORGANIC MATTER) 100	165
	(CARBONATE ION, EROSION)	500			(CALCIUM CARBONATE) 500	
CALCIUM	(EROSION)	500	5,000	25	(CALCIUM CARBONATE) 500	1,000

RATIOS of carbon and calcium atoms to phosphorus atoms are given for inputs to the sea, average sea salt, marine organisms and marine sediments, along with the average length of time atoms of each element reside in the ocean before being buried in sediment. A distinction is made between carbon added by volcanism and added by erosion of calcium carbonate ($CaCO_3$) and between carbon buried in sediments as organic matter and buried as calcium carbonate.

	WARM SURFACE WATER	AVERAGE DEEP WATER
PHOSPHATE	0	2.2
TOTAL DISSOLVED CARBON	2,000	2,275
BICARBONATE ION	1,764	2,140
CARBONATE ION	223	90
CARBON DIOXIDE GAS	13	45
CALCIUM ION	10,000	10,055
CADMIUM	0	.7
$\frac{\text{CARBON 13}}{\text{CARBON 12}}$ (COMPARED TO STANDARD RATIO)	1.0024	1.0002

COMPOSITION of warm surface water is compared with that of average deep water. The concentration of the first six constituents is given in millionths of a mole per liter and that of cadmium in billionths. The ratios of carbon 13 to carbon 12 are compared with the standard ratio, in a National Bureau of Standards reference calcium carbonate. The differences between surface- and deep-water values observed for both chemical and carbon-isotope composition arise because organisms take up constituents in surface water; their soft-tissue residues and calcite ($CaCO_3$) shells fall into the deep sea, where decomposition and dissolution return the constituents to solution. A balance of electric charge is maintained: the sum of the charges contributed by positive calcium ions and by negative bicarbonate and carbonate ions remains constant.





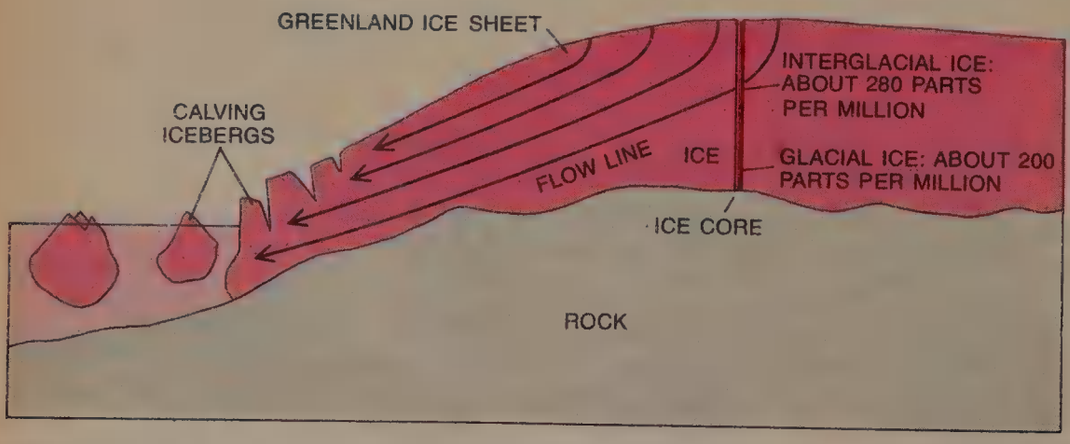
CARBONATE CONTENT of ocean water decreases with depth (*colored curves*), whereas the saturation content increases with depth (*gray curve*). The level where the two curves intersect is the lysocline. Water is supersaturated with the mineral calcite above the lysocline, undersaturated below it. There is evidence that after the last glacial era the lysocline was deeper than it had been before and than it is now. The carbonate content of deep water must have been greater (*broken line*).

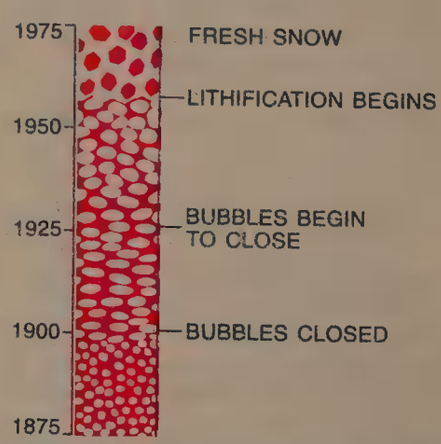
LYSOCLINE separates the ocean's sediments into two broad categories. Sediments below the lysocline, where the water is undersaturated with respect to calcite, have lost their calcite to dissolution, and so they are dominated by clay minerals eroded from the continents: the red clays. Sediments above the lysocline, laid down in water that is supersaturated with respect to calcite, are dominated by the undissolved calcite chambers that were formed by plankton. These sediments are often speckled with the shells of foraminifera, and so they are called foram oozes. The lysocline moves in response to changes in the chemical composition of sea salt, always seeking a level that allows the loss of carbon from ocean water to match the rate of input of carbon to the ocean. In today's ocean marine organisms are producing calcite at a rate that takes up carbon faster than it is supplied by erosion and volcanism. The ocean's chemistry has adjusted itself so that this overproduction of calcite is compensated for by the dissolution of calcite.

Above the lysocline, the depth at which the two curves cross each other, the water is supersaturated with calcite; below it the water is undersaturated. As a result most of the calcite raining onto sea-floor ridges and plateaus above the lysocline is preserved; its carbonate is buried and removed from circulation. Below the lysocline most of the calcite is dissolved and its carbonate is returned to circulation.

Because surface water is supersaturated with calcium carbonate, organisms can manufacture as much calcite as they require. In today's ocean, organisms annually take up several times as much carbon to form calcite as is supplied by erosion and volcanism. The overproduction of calcite is compensated for by changes in the ocean's chemistry that keep the burial of calcium carbonate in balance with the supply of carbon. The lysocline moves up and down in response to the changing concentration of carbonate ion, seeking the level at which the amount of calcite that dissolves just matches the overproduction of calcite by organisms. If, for example, the burial rate of calcite were temporarily to become higher than the rate of supply of carbon destined to be incorporated in calcite, the ocean's carbonate content would begin to decrease. This would lead to a rise in the lysocline and a consequent enlargement of the area of the sea floor over which calcite is dissolved instead of being buried. The imbalance between loss and gain would gradually be restored.

The control systems I have been describing make it possible to calculate the concentrations in the ocean of phosphate and carbonate for any given set of environmental conditions. To quantify the bicarbonate (HCO_3^-) content one turns to charge balance. Discounting a host of minor ions, one can say that





ICE CORES supply evidence for about a 40 percent increase in atmospheric carbon dioxide at the end of the last glacial episode. One deep borehole into the Greenland ice sheet penetrated ice that had accumulated for 80,000 years (*left*). The core contains bubbles of air that were trapped as snow was lithified to become solid ice; such bubbles have been shown to close, and thus seal their air sample, some 50 to 75 years after the snow falls (*right*). In the laboratory of Hans Oeschger of the University of Bern one-centimeter cubes are cut from cores and crushed in a vacuum. The total pressure of the released air is measured and the carbon dioxide content is determined with a laser device. Air trapped in ice some 18,000 years ago, during glacial time, was found to have about 200 parts per million of carbon dioxide; air trapped in the recent past had about 280 parts per million. (Carbon dioxide level has now risen to about 340 parts per million.)



Photo: courtesy of NASA.

ON SOME TRIPS, YOU CAN'T JUST TOSS YOUR LUGGAGE IN THE TRUNK.

When you're sending important cargo into space, you have to pack carefully.

It's called space cargo processing and McDonnell Douglas does it.

Space cargo is precision machines, laboratories, instruments, factories, satellites, rockets and experiments. Space cargo is intricate, delicate and expensive. It must be planned for, integrated, and guarded. Some must be refrigerated; some must be heated. Some is put in place with tweezers and some with cranes. All must be packaged to fit the cargo hold of NASA's shuttles, attached to their operating controls and carefully fitted to withstand the forces of the journey.

And when the journey's over, the results must be harvested. Unloaded, preserved, packaged and delivered.

We've been sending spacecraft and launch systems into space for thirty-four years—that's why we know that when going to work in space, it's important how you pack.

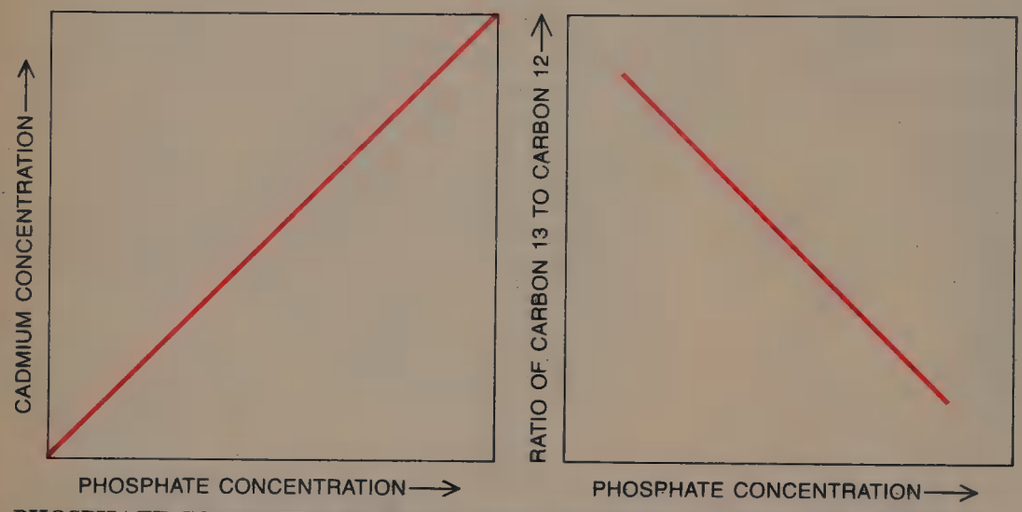
MCDONNELL DOUGLAS

the total concentration of five positively charged ions (sodium, potassium, magnesium, calcium and hydrogen) must equal that of five negatively charged ions (chloride, sulfate, bicarbonate, carbonate and hydroxyl). The hydrogen-ion and hydroxyl-ion concentrations are small enough in seawater to be neglected. On the time scale of glacial episodes on the earth the concentrations of all the other species except carbonate and bicarbonate remain constant. The carbonate ion's concentration has been established. Only the bicarbonate ion remains. Its concentration adjusts to a value corresponding to electrical neutrality, that is, it is equal to the sum of the concentrations of the positive and the negative ions listed above.

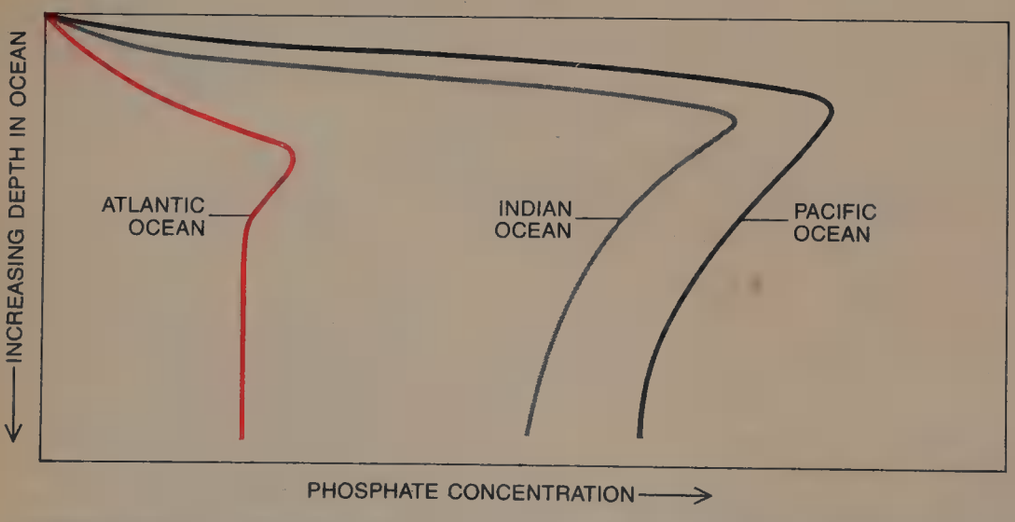
At the depth of the lysocline the carbonate-ion content is at the saturation level, and so there the bicarbonate-ion content is fixed by the charge-balance requirement. Having fixed the lysocline-depth composition, one can calculate that of surface water by removing phosphate and the associated carbon dioxide (for soft tissue) and carbonate (for calcite) from the lysocline composition. Now, having fixed the carbonate and bicarbonate content of surface seawater, one has also fixed its carbon dioxide content: since bicarbonate is formed from (and dissociates into) carbon dioxide, carbonate and water, the surface-water concentration of carbon dioxide is equal to the ratio of the square of the bicarbonate-ion concentration to the carbonate-ion concentration (multiplied by a constant that varies with temperature and salinity).

I have not yet mentioned the control system for dissolved oxygen (O_2). The surface water's dissolved-oxygen content is dictated by the solubility of the gas. The deep-water content is set by the phosphate concentration of the deep sea. When deep water reaches the surface, its phosphorus atoms are all converted into (and ultimately lost as) organic residues. The residues are largely consumed by organisms living in the deep sea, creating a demand on the dissolved oxygen there. Overall about half of the oxygen reaching the deep ocean with newly sunken bottom water is consumed by animals and bacteria. The supply depends on the ratio of the amount of oxygen in the air (and hence the amount dissolved in cold surface water) to the amount of phosphorus dissolved in the ocean. Since on the time scale of glacial episodes the oxygen content of the atmosphere cannot change significantly, the dissolved-oxygen content of deep water varies inversely with the amount of dissolved phosphorus.

The dependence of oxygen content on phosphate content reinforces the feedback mechanism tending to balance phosphate loss and gain. In discussing that control mechanism I made the assumption that the fraction of the residues falling into the deep sea and surviving dissolution remains constant even as the water's phosphate content varies. Actually the probability of dissolution is influenced by the degree of oxygenation of the deep water. Since animal life demands oxygen, the efficiency of the scavenging community drops as the oxygen content of the water decreases. Where there is no oxygen there is no animal activity, and the residues are consumed only by the less efficient community of anaerobic bacteria. (In isolated basins where there is no dissolved oxygen at all the fraction of organic residues that are preserved in sediments is observed to increase substantially.) In situations where phosphorus is leaving the ocean faster than it is being supplied by erosion the phosphorus content drops and the deep-water dissolved-oxygen content rises. The decrease in phosphorus reduces the rate of formation of organic residues; the increase in oxygen reduces the fraction of those residues that escape oxidation and are buried. The two feedback mechanisms work in harmony to reduce the rate of loss of phosphorus to sediments, thereby forcing the system back toward balance.

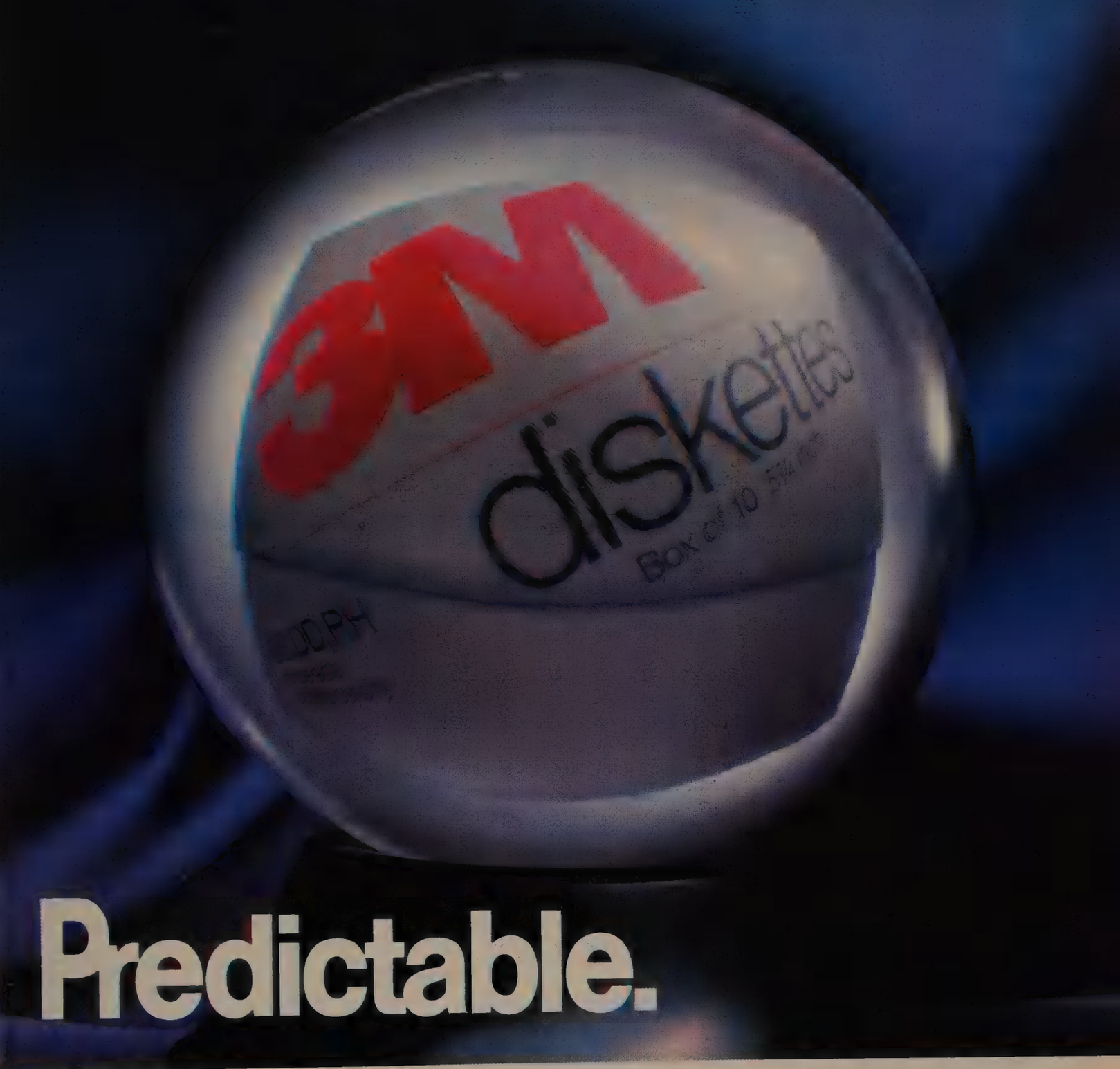


PHOSPHATE CONTENT of the ocean is reflected by two markers in the shells of fossil foraminifera: the content of cadmium and the ratio of the two stable isotopes of carbon. Cadmium is a marker because its concentration varies directly with that of phosphate in seawater (*left*); the ratio of the rarer carbon 13 to the much more abundant carbon 12 in the ocean's total dissolved carbon is a marker because it varies inversely with the concentration of phosphate in seawater (*right*). Analysis of foraminifera from different periods indicates that there was a decrease in the ocean's overall phosphate content at the close of the last glacial period.



DISTRIBUTION OF PHOSPHATE in the ocean varies with depth. Essentially all the phosphate in surface water is taken up by marine plants and falls into the deep sea in soft-tissue residues. Much of it is released into solution as the residues are decomposed by animals and bacteria, a process that takes place primarily at intermediate depths. Superimposed on this vertical variation there is an Atlantic-to-Pacific increase in the deep-water phosphate content. It results from the flow of newly sunken deep water, much of which now originates in the northern Atlantic and tends to purge the Atlantic of the products of decomposition. The horizontal gradient complicates attempts to reconstruct the chemical composition of the glacial ocean: climatic events may have altered the deep-current pattern as well as the ocean's bulk chemistry.

Knowing something about the mechanisms controlling ocean chemistry, one is in a position to evaluate evidence for changes in the past and perhaps to predict changes in the future. What hard



Predictable.

A flawless future is in sight with 3M diskettes.

When it comes to keeping track of precious data, predictable means reliable. Being able to count on every diskette, every time. At 3M, reliability is built into every diskette. We've been in the computer media business for over 30 years. And we've never settled in. We're constantly improving and perfecting our product line, from computer tape and data cartridges to floppy disks.

3M diskettes are made at 3M. That way, we have complete control over the entire manufacturing process. And you can have complete confidence in the reliability of every 3M diskette you buy.

Look in the Yellow Pages under Computer Supplies and Parts for the 3M distributor nearest you. In Canada, write 3M Canada, Inc., London, Ontario. If it's worth remembering, it's worth 3M diskettes.



3M hears you...

3M

evidence is there that ocean chemistry has indeed varied over time? The most spectacular and convincing source of information comes from ice cores removed by deep boring into the Greenland and Antarctic ice sheets. The cores reach down to ice that was formed 80,000 years ago. By crushing samples of the ice in a vacuum it is possible to release what is in effect fossil air, trapped in tiny bubbles as new-fallen snow consolidated to become solid ice.

The carbon dioxide content of the released air has been measured, in samples going back 40,000 years, by investigators at the University of Bern and at the University of Grenoble. Both groups got the same striking result: air from ice formed during the peak of the most recent glacial episode, some 18,000 years ago, has only about 70 percent as much carbon dioxide as air from ice formed during the present interglacial period.

An increase of this magnitude in atmospheric carbon dioxide could have been caused by a rise of about 12 degrees Celsius in the temperature of surface ocean water. From rather good independent evidence it is known, however, that the postglacial rise in water temperature was only about two degrees, and the small increase in atmospheric carbon dioxide generated by that amount of warming would have been largely countered by the effect of the oceanwide decrease in salinity resulting from the melting of glaciers. The temperature explanation can apparently be ruled out. The remarkable increase in atmospheric carbon dioxide at the end of glaciation will have to be explained by changes in ocean chemistry.

Another indication of a chemical change on a glacial time scale is a curious large and rather brief deepening of the ocean's lysocline early in the postglacial period. The evidence is in data showing the depth at which the shells of certain marine animals, which would have dissolved had they been below the lysocline, were instead preserved in sediments. The implication is that the concentration of carbonate in the deep sea increased by about 10 or 20 percent at the close of glacial time some 11,000 years ago, reaching a maximum early in postglacial time and then dropping back to about its glacial value perhaps 5,000 years ago. If that is the case, the ocean's calcium carbonate budget must have been seriously perturbed by events associated with deglaciation.

Finally, there are two other items of evidence, which can be interpreted to indicate a substantial decrease in the ocean's phosphate content at the close of the last glacial period. Both come from analysis of the shells of the unicellular marine animals called foraminifera. One finding is a decrease in the content of the element cadmium in the shells of bottom-dwelling foraminifera. The other has to do with the ratio of the two stable isotopes of carbon, carbon 13 and carbon 12, in the shells of foraminifera.

Both cadmium and the carbon-isotope ratio serve as markers for the ocean's phosphate content. The concentrations of dissolved cadmium and phosphate in the sea are closely correlated: both go to zero in the warm surface water, indicating that plants take them up efficiently, and they must have the same pattern of release from organisms deeper in the ocean. Similarly, the ratio of dissolved carbon 13 to carbon 12 is correlated (but inversely) with the phosphate concentration: the higher the phosphate content, the lower the ratio. The reason is clear. In photosynthesis plants take up the abundant lighter isotope, carbon 12, in slight preference to the much less abundant carbon 13. The dissolved carbon of surface waters (where there is almost no phosphate) is therefore slightly enriched in the heavy isotope: its 13-to-12 ratio is higher than the oceanwide average. The organic residues that fall into the deep sea, on the other hand, are eventually oxidized, giving up their excess carbon 12 to the ambient water and thereby lowering the 13-to-12 ratio of the dissolved carbon in



WITHDRAWAL OF GLACIERS some 11,000 years ago was accompanied by a rise in the sea level, which flooded what had been exposed continental margins, and also by a proliferation of forests. The maps, developed by George J. Kukla of the Lamont-Doherty Geological

deep water—where the phosphate content is high.

Edward A. Boyle of the Massachusetts Institute of Technology demonstrated that the cadmium content of foraminifera in recent sediments varies linearly with the cadmium content of the water in which they grew; the content of the element in bottom-dwelling foraminifera shells from deep-sea cores should therefore reflect changes in its concentration in bottom water over time. In cores from various depths in the North Atlantic, Boyle has found a dramatic twofold decrease in foraminiferan cadmium content at the close of glacial time. The implication is that the phosphate content dropped significantly then. How much of the drop resulted from changes in the chemistry of the entire ocean and how much from changes in the pattern of deep-water circulation is not yet known; measurements in cores from other regions should make that clear. At this early stage, however, Boyle's finding provides at least a qualitative indication that there was more phosphorus in the deep sea during glacial time than there was after the end of glaciation.

Examination of sediment cores by Nicholas J. Shackleton of the University of Cambridge has shown that there was an abrupt increase in the ratio of carbon 13 to carbon 12 in the shells of bottom-dwelling foraminifera at the close of glacial time. No such change is observed in the shells of surface-dwelling foraminifera. Because the volume of the deep-water reservoir is much larger than that of the surface reservoir, the mean 13-to-12 ratio of the carbon in the ocean as a whole (and in the atmosphere) must have increased at the close of glacial time. The only reasonable explanation seems to be a net loss of carbon dioxide from the ocean-atmosphere system to organic matter, which would have taken up carbon 12 preferentially; the carbon left in the atmosphere and left dissolved in the ocean would have been slightly enriched in carbon 13. The manufacture of organic carbon could have taken place in the forests that proliferated with the withdrawal of the glaciers. I shall have more to say about this possibility below.

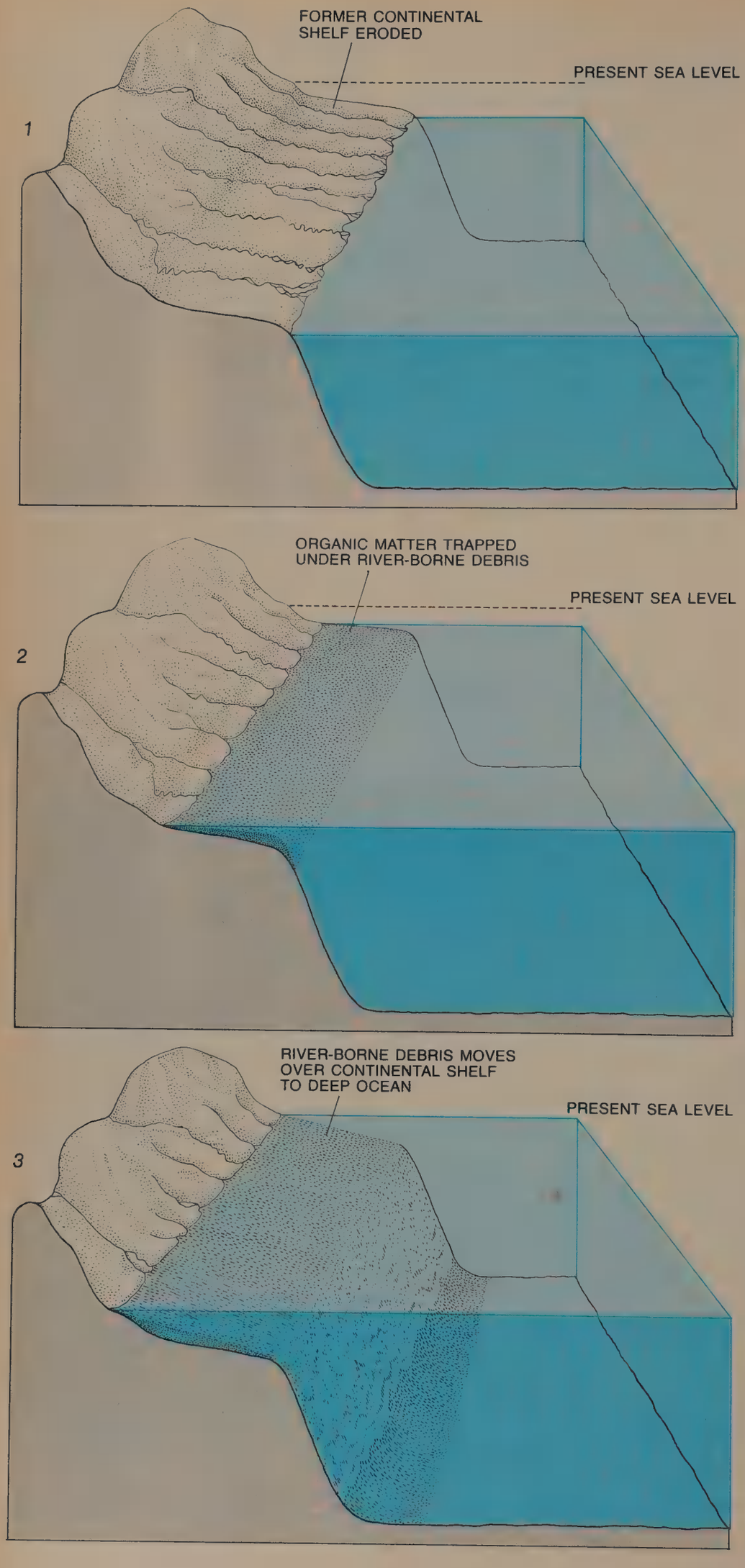
The alternative is that organic carbon manufactured by marine plants in coastal waters was sequestered by burial and storage in shallow sediments, removing carbon—and preferentially carbon 12—from the ocean. Such sediments would have formed on the continental shelves as the glacial ice sheets melted, causing the sea level to rise and transgress the land. The shelf-storage hypothesis, which assumes that the oceanwide change in the carbon-isotope ratio resulted from the storage of organic residues in continental-shelf sediments, can account for all the other chemical changes I have mentioned, provided that cadmium, phosphorus and carbon were present in roughly the same ratios in the trapped residues as they are in residues now falling from the surface into the deep sea.

Even the fact that no postglacial change is observed in the carbon-isotope ratio of surface-dwelling foraminifera can be explained by shelf storage. Because the ratio of phosphorus to carbon in organic residues is about 10 times as high as the ratio in seawater, the sequestering of organic residues removes relatively more phosphorus than carbon from seawater, and so the ratio of phosphorus to carbon would have been reduced by the shelf-storage event. Since phosphorus is the limiting agent for life, the reduction in the ratio would have reduced the extent to which upwelling oceanic carbon was incorporated into organic matter. This in turn would have caused the increase in the carbon dioxide content of surface water, and hence of the atmosphere, that is recorded in ice cores. Another result would have been a decrease in the ratio of carbon 13 to carbon 12 in surface water, which would have canceled the oceanwide increase recorded in bottom-dwelling foraminifera. Still another result would have been



Observatory of Columbia University, contrast conditions at the peak of the glacial era (*left*) with conditions now (*right*). (The current map shows "potential" forestation in the absence of human interference.)

One hypothesis to explain chemical changes at the end of the glacial era depends on the burial of organic matter on newly submerged continental shelves. Another depends on the proliferation of forests.



the reduction of the cadmium content of ocean water.

The temporary deepening of the lysocline in early postglacial time could also have resulted from shelf storage. The uptake of carbon by organic matter that was subsequently buried would have left less carbon in the ocean to form the negative ions carbonate and bicarbonate. The charge demand of the ocean's excess positive ions would have forced a larger proportion of the carbon ions to be in the doubly charged state: carbonate rather than bicarbonate. The increase in carbonate would have deepened the isocline. A deeper isocline, however, means more deposition of undissolved calcium carbonate formed by marine organisms, and so the excess carbonate content would gradually have been reduced. In several thousand years the isocline would have returned to about its previous level, just as the fossil record indicates.

I have mentioned the possibility that the expansion of forests at the close of glacial time could have accounted for the reduction of the carbon-isotope ratio. The trouble is that the forests would have taken up from the ocean-atmosphere system only carbon dioxide and not phosphorus. If the growth of new forests had been the only perturbation, the carbon dioxide content of the air would have decreased, whereas the ice cores say it increased.

There is, however, a scenario according to which a combination of forest expansion and ecological shifts in the ocean could generate all the observed postglacial changes. In order for the scenario to work the ratio of phosphorus to carbon in residues falling into the deep sea would have to have been about 30 percent lower during glacial time than it is today. If marine organisms incorporated more carbon in relation to phosphorus in their soft tissue during glacial time and relatively less carbon after glaciation, the ocean's content of carbon dioxide (and therefore the atmosphere's

SHELF-DEPOSITION HYPOTHESIS cites this sequence of events to account for changes in ocean and atmospheric chemistry. During full glacial time (*1*) the sea level was about 150 meters below its present level. What had been (and is now) the continental shelf was exposed and underwent erosion. During the transition to interglacial time (*2*) the sea rose, flooding the sculptured shelf and river estuaries. Organic matter from the coastal ocean was trapped as river-borne debris rebuilt the eroded margin, and the burial of organic matter reduced the carbon and phosphorus content of sea salt. After the sea level stabilized some 6,000 years ago (*3*) the continental margins reached a steady-state shape, allowing river debris to be carried over the shelf into deeper water; the rapid drain of carbon and phosphorus from the ocean came to an end.

carbon dioxide content) would indeed have increased. There is now no way of saying whether or not such a large change in the phosphorus-to-carbon ratio in organisms could have taken place.

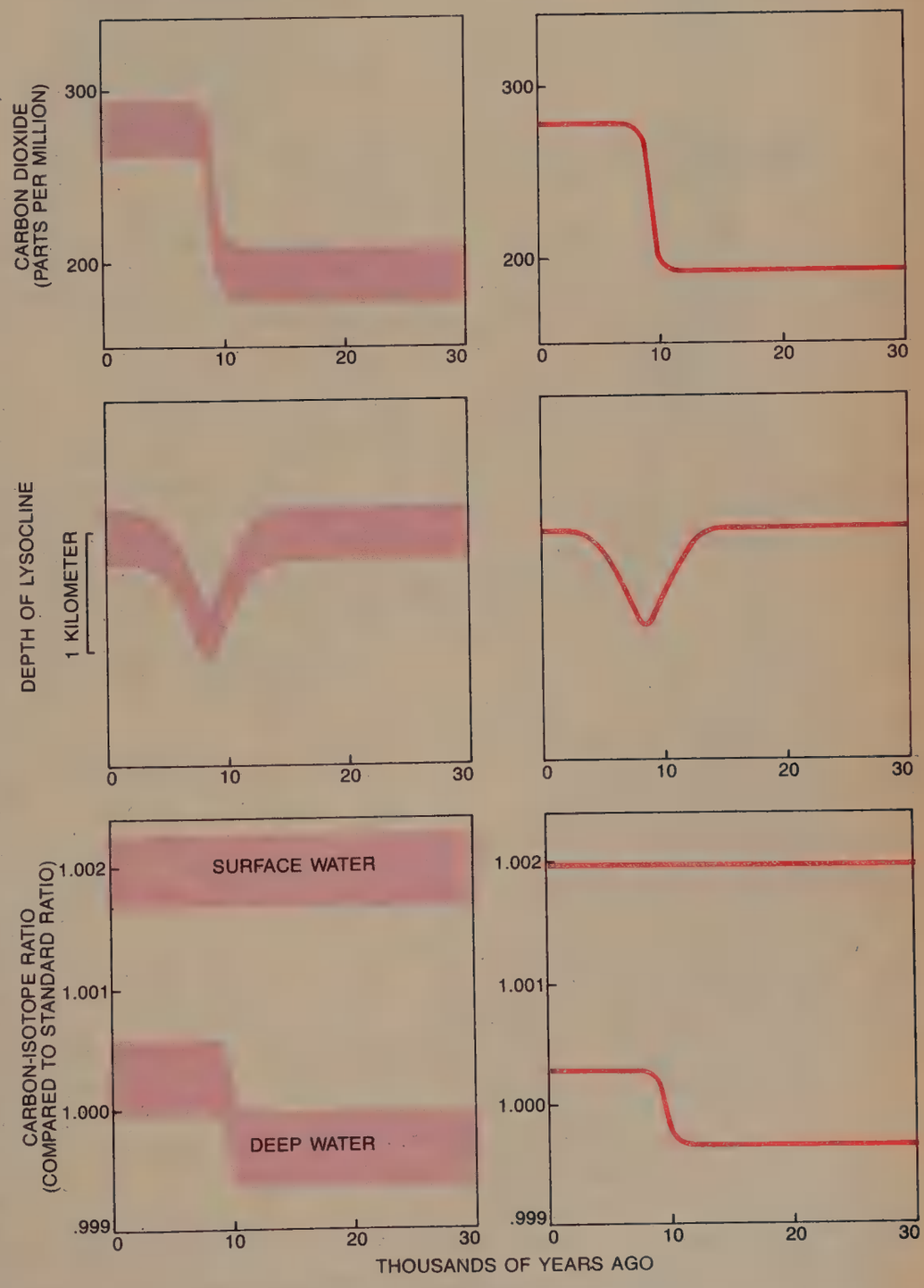
Both the shelf-storage hypothesis and the one based on forest growth and the phosphorus-to-carbon ratio can explain most of the observed postglacial changes, and so one cannot yet choose between them. Fortunately the two mechanisms would have had very different effects on the ocean's phosphorus content, and therefore also on its cadmium. The shelf-deposition hypothesis predicts a change in the ocean's cadmium content, whereas the residue-ratio hypothesis predicts no such change. The measurement of cadmium in bottom-dwelling foraminifera from the Indian Ocean and the Pacific, which is now in progress in Boyle's laboratory, should point to one hypothesis or the other.

What is currently being learned about the chemical changes in the ocean at the end of the most recent glaciation may make it possible to generalize somewhat about chemical changes associated with other glacial cycles. It is impossible, however, to draw up a scenario going back beyond the series of major glacial cycles that have dominated the past million years of the earth's history. For one thing the carbon dioxide record in ice now goes back only 40,000 years; even if data could be extracted from older samples, the oldest ice on the earth (deep in the middle of the Antarctic ice sheet) is probably no more than half a million years old. Moreover, the number of variables would keep increasing; for example, neither the concentration of calcium ions in the sea nor the atmosphere's oxygen content could be assumed to remain constant. On even longer time scales, say tens of millions of years, such factors as rates of erosion and rates of deep-sea oxygenation must have been different from what they are now because the basic geometry of the ocean basins and continents was different. New and powerful indicators of paleochemistry will have to be discovered before changes in ocean chemistry can be read reliably over these very long time scales.

This is not to say that isolated, dramatic changes in ocean chemistry may not be detectable if they attended catastrophic events that left clear evidence in ancient fossil records. Luis W. Alvarez, Walter S. Alvarez and their colleagues at the University of California at Berkeley have shown that the profound extinctions marking the close of Cretaceous time 60 million years ago were quite likely triggered by the impact of a large extraterrestrial object, either an asteroid or a comet. The sedimentary record reveals that the event took a particularly large toll on marine life: more than half of the Cretaceous species suddenly disappeared.

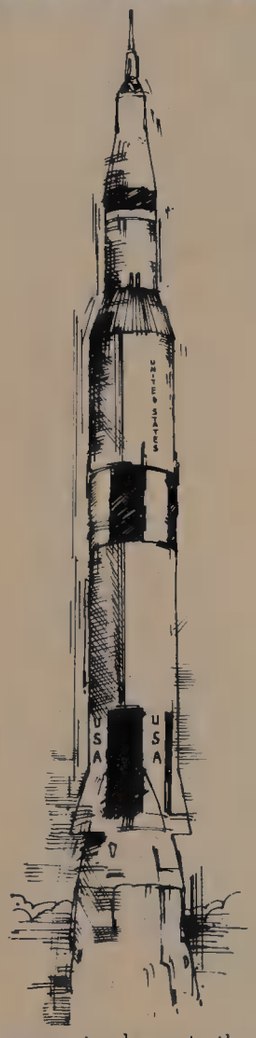
If the shock to the web of life was strong enough to have significantly reduced the efficiency with which upwelling phosphorus was incorporated by surviving marine plants, the chemistry of the surface ocean would have changed dramatically, becoming more like that of the deep sea. By analogy with today's ocean one can imagine such a change would have caused a sharp rise in atmospheric carbon dioxide. If the reduction in plant life continued long enough, it would be recorded as a change in the ratio of carbon 13 to carbon 12 in the calcite formed by surviving planktonic organisms. The current worldwide search for evidence bearing on the event that ended the Cretaceous period may turn up such a carbon-isotope anomaly. Even if it does not, it should still yield valuable information about the sequence of events immediately following the impact disaster, and perhaps some clues to chemical changes associated with other such disasters.

In this article I have treated variations in ocean and atmospheric chemistry as the result of chance perturbations in factors controlling the flow of material through the sea to its sediments. J. E.



GEOLOGIC RECORD gives observed values (*left*) for the atmospheric content of carbon dioxide (*top*), the depth of the lysocline (*middle*) and the carbon-isotope ratio (*bottom*). The range of observed values brackets the values (*right*) calculated when a two-reservoir ocean is assumed to have been perturbed (by the removal of about 3 percent of the carbon that was present in glacial time and about 30 percent of the phosphorus) as a result of shelf deposition.

Technology is in a state.



Tennessee is closer to the moon than it looks: much of the technology launched into space was researched in Tennessee. But space is just one side of Tennessee's technology story. Electronics and computer science are major industries. The University of Tennessee is becoming a nationally recognized energy research center, and we're making important breakthroughs in medical technology. When it comes to high technology, Tennessee is getting down to brass tacks and giving you the moon, and more. For more information, write or call Mike DuBois, Tennessee Department of Economic and Community Development, Andrew Jackson Bldg., Box 726, Nashville, TN 37219. 1-800-251-8594.

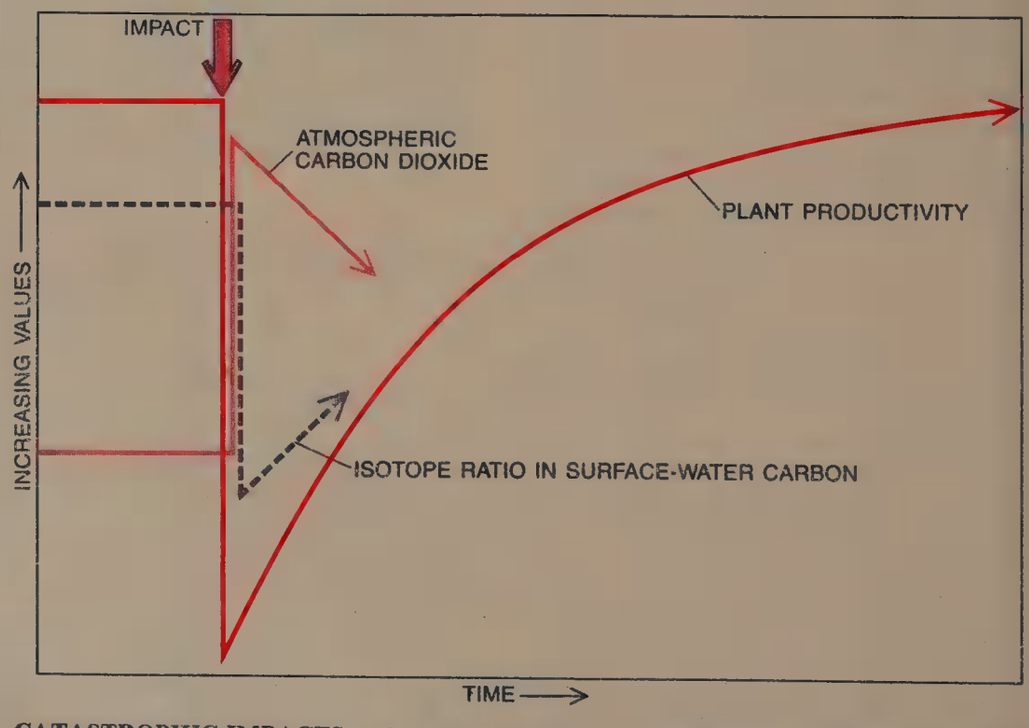
Tennessee

We get down to Brass Tacks.

	SHELF-DEPOSITION HYPOTHESIS	FOREST AND RESIDUE-RATIO HYPOTHESIS	OBSERVED
PHOSPHATE	DOWN 30 PERCENT	NO CHANGE	NO EVIDENCE
CADMIUM	DOWN 30 PERCENT	NO CHANGE	DOWN 50 PERCENT (IN DEEP ATLANTIC)
CARBON 13 / CARBON 12 (MEAN OCEAN)	UP .07 PERCENT	UP .07 PERCENT	UP ABOUT .06 PERCENT
CARBON 13 / CARBON 12 (SURFACE OCEAN)	NO CHANGE	NO CHANGE	NO CHANGE
LYSOCLINE DEPTH	DOWN 1 KILOMETER	DOWN 1 KILOMETER	DOWN 1-2 KILOMETERS
TOTAL DISSOLVED CARBON	DOWN 3 PERCENT	DOWN 3 PERCENT	NO EVIDENCE
ATMOSPHERIC CARBON DIOXIDE CONTENT	UP 40 PERCENT	UP 40 PERCENT	UP 40 PERCENT

TWO HYPOTHESES, one based on shelf deposition and the other on a combination of forest growth with an altered composition of organic residues falling into the deep sea, explain the observed changes in carbon-isotope ratio, depth of the lysocline and atmospheric carbon dioxide content. They predict different changes in phosphate (and so in cadmium) content, however.

Lovelock, in his book *Gaia: A New Look at Life on Earth*, raises the possibility that the community of living things has not only interacted with the inorganic world but has somehow orchestrated the chemistry of the atmosphere to its own advantage. Perhaps. Certainly chemical change has played a major role in the development of life on the planet. The ocean-atmosphere system has had to adjust its chemistry over rather a wide range to maintain a balance between the loss and gain of many of the constituents of sea salt, and organisms have effected some of the changes as well as being affected by them.

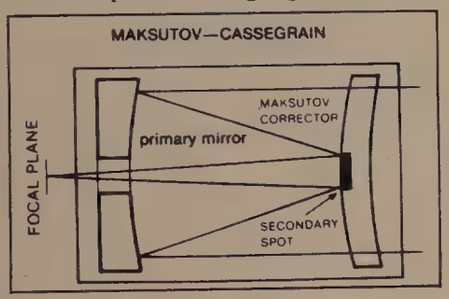


CATASTROPHIC IMPACTS such as the one thought to have ended the Cretaceous period might have generated changes in ocean chemistry that are recorded in the fossil record. If the effects of an impact temporarily devastated marine plant life, there would have been an abrupt increase in surface-water carbon dioxide and in atmospheric carbon dioxide. Eventually plant productivity would have been restored, reducing the carbon dioxide level. Changes in plant productivity should be recorded as changes in the carbon-isotope ratio in foraminiferan shells.

NEWTON'S ILLUSION, CELESTRON'S® C-90—AND YOU.

In the beginning, all telescopes were refractors: the larger their objective lens, the longer their tube and the greater their expense. The newtonian reflector sent light down the tube to a parabolic main mirror, which reflected it back to the eyepiece. The tube reduced weight and cost and gave a brilliant image. Newton's mirror revolutionized optical design.

Today, computers, laser technology and better optics allow the revolution to continue. Witness the Celestron C-90: using precision mirrors to "fold" the image, it allows forty inches of focal length in a tube about eight inches long! Its advanced MAKSUTOV-CASSEGRAIN optical design gathers 159 times more light than the human eye. And, it weighs less than 9 pounds. A refractor of the same size would weigh 20 times as much, a newtonian, 10 times as much.





Because of its precision optics and portability, the Celestron C-90 has become one of the most popular instruments in a line of world class telescopes. Many purchase a C-90 to test their interest before moving up to one of our larger telescopes, such as the C-8, the world's most popular modern telescope.

The C-90 can be equipped for use as a telescope, spotting scope or telephoto. All fully capable of visual and photographic use. Each comes complete with Celestron Certified Optics and a 25 year limited warranty.

The perfect, personal, portable telescope is at your nearest Celestron Dealer. Call for more information regarding the C-90, and Celestron's other telescopes, binoculars, spotting scopes and telephotos.

CELESTRON®
PRECISION OPTICS

For Dealer Locations, Catalogs, Information, Technical Data Sheet

CALL FREE
1-800-421-1526
(Mon-Fri: 8-4:00 Pacific Time; Continental USA only)

CELESTRON INTERNATIONAL, P.O. Box 3578SA9,
2835 Columbia St., Torrance, CA 90503 (213) 328-9560

Noi parliamo il linguaggio della gente che fa il futuro.

LE SCIENZE is the Italian edition of SCIENTIFIC AMERICAN. It's read by 93,900* Italian professionals working in business, industry and government.

The people responsible for Italy's achievements on the technology front are reading LE SCIENZE. The highly educated and well-paid professionals that read the Italian edition of SCIENTIFIC AMERICAN can afford a comfortable and good way of life. They represent the brains that will make Italy's future happen and the power to make decisions for that future.

We speak the language of more than 3,000,000 world shakers in 140 countries each month. We reach the people who make the future happen in Italian, French, Spanish, German, Japanese, Chinese and English.

For more information on SCIENTIFIC AMERICAN call:
Milan: (011-39) 2-805-9874
London: (011-441) 839-2873
New York: (212) 754-0262

LE SCIENZE
edizione italiana di SCIENTIFIC AMERICAN



*ADS audited

The Atmosphere

Its dynamic activity serves to distribute the energy of solar radiation received by the earth. Models of this activity help to explain climates of the past and predict those of the future

by Andrew P. Ingersoll

The atmosphere is the working fluid of the earth's heat engine. Most of the radiant energy arriving from the sun is converted into atmospheric heat energy before it is reradiated into space. The winds redistribute this energy, dissipating more of it in the process than is dissipated by ocean currents, tides, continental drift and mantle convection combined. Both the short-term fluctuations of the atmospheric system (the weather) and the longer-term fluctuations of the average weather (the climate) are an important part of earth history.

The composition of the atmosphere bears little relation to that of the cloud of gas and dust out of which the earth formed. Rather, the composition of the air reflects a complex history of the reactions of volatile elements with the dust in the primordial solar nebula, with magma rising from the earth's mantle, with the rocks of the earth's crust, with the ocean and with the biosphere.

The three major constituents of dry air are nitrogen (N_2), oxygen (O_2) and argon (Ar), which account respectively for 79 percent, 20 percent and 1 percent of the molecules. Nitrogen is for all practical purposes geochemically inert and has therefore accumulated in the atmosphere. Oxygen, on the other hand, cycles through the atmosphere and also through the ocean, the biosphere and sedimentary rocks. The amount of oxygen in the atmosphere is set by the rate of reactions that link the atmospheric reservoir of free oxygen to the reservoir of reduced carbon in sedimentary rocks. The small fraction of organic matter buried in sediments before it can decay corresponds to a net addition of oxygen to the atmosphere; the rate at which oxygen is produced by this means is balanced by the rate at which it is consumed by the weathering of sedimentary rocks. Most of the argon in the atmosphere is the isotope argon 40, produced by the radioactive decay of potassium 40 in the mantle and the crust and released into the atmosphere by volcanoes. Argon, one of the noble gases, is geochemically inert; once it is released into the atmosphere it stays there.

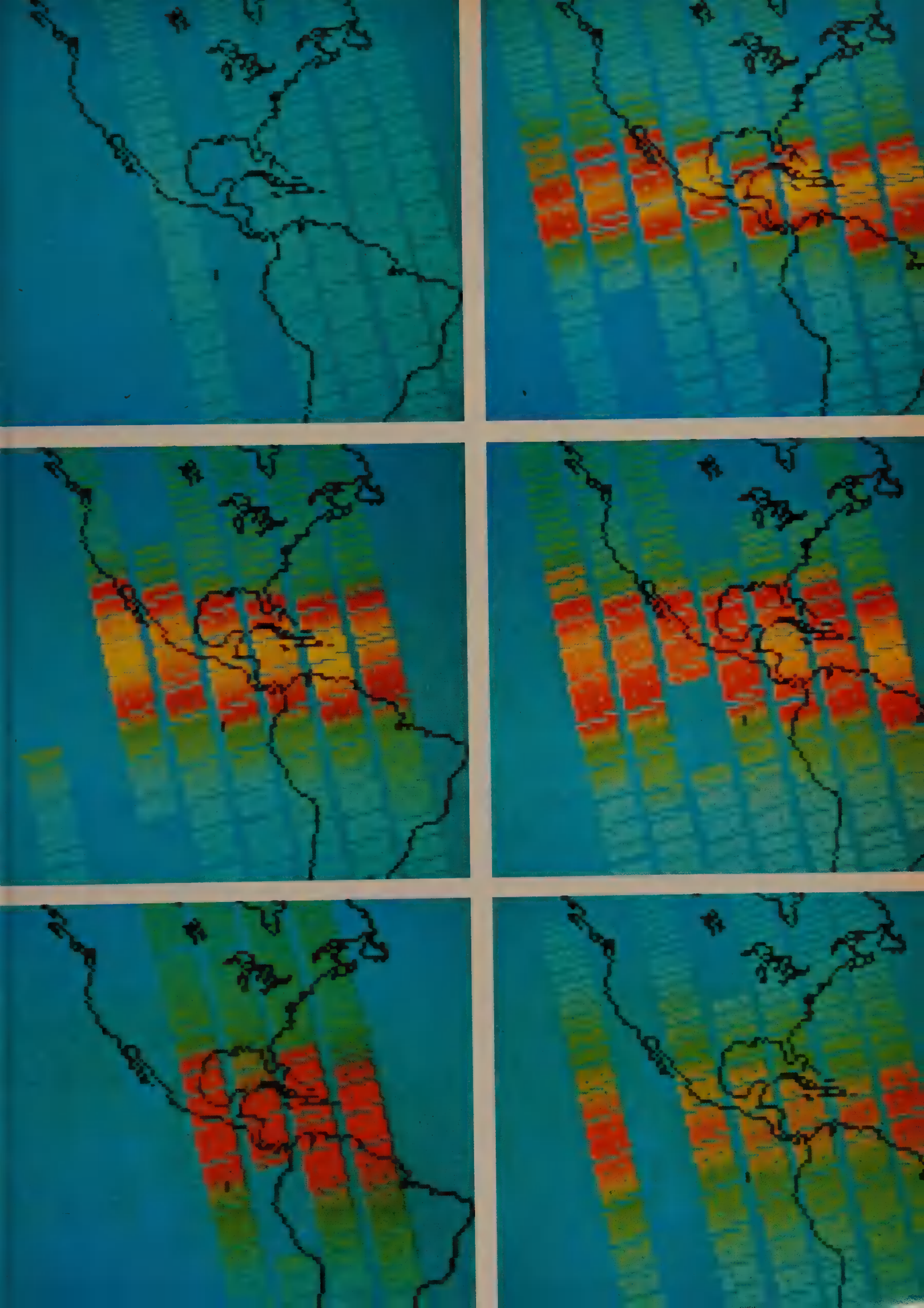
The composition of dry air is remarkably constant over the globe but the amount of water vapor in the atmosphere varies widely, ranging from 4 percent by volume to a small fraction of 1 percent. The atmosphere is not the dominant reservoir of water because water condenses at average earth temperatures. The ocean holds some 300 atmospheres of water (that is, enough water to equal 300 times the mass of all the constituents of the atmosphere), and clays and other hydrous minerals hold a somewhat smaller amount. The water-vapor content of a given air mass is therefore determined by the history of its contact with the surface reservoirs of water and by the saturation vapor pressure of air at different temperatures.

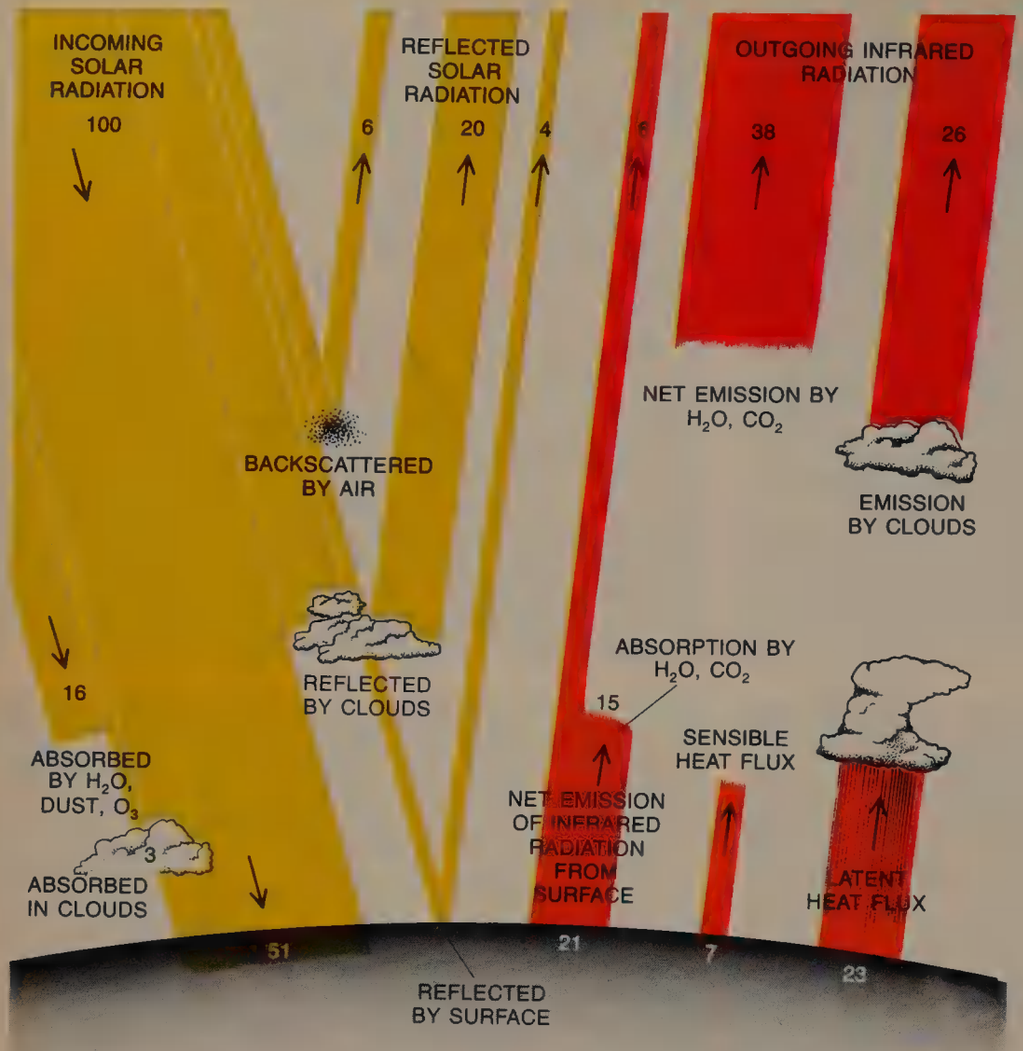
The remaining constituents of air are present in such small amounts that their concentrations are generally given in parts per million rather than in parts per 100 (percent). The most abundant of them is carbon dioxide (CO_2), which at present accounts for about 340 parts per million of dry air. The atmospheric reservoir of carbon dioxide is dominated by the much larger oceanic reservoir of bicarbonate (HCO_3^-) and carbonate (CO_3^{--}) ions and by the reservoir of carbon in limestone rocks, consisting largely of the calcium carbonate ($CaCO_3$) shells of marine organisms. The limestone rocks alone contain some 20 atmospheres of carbon dioxide. The amount of carbon dioxide in the air, like that of water vapor, is therefore set by the rates of the equilibrium reactions linking the atmospheric reservoir to surface reservoirs.

The most abundant gases after carbon dioxide are neon (18 parts per million) and helium (five parts per million). Most of the helium is produced by the decay of radioactive elements in the solid earth. The neon, however, is primordial. The amounts of neon and of the other rare gases (krypton, xenon and the nonradiogenic isotopes of argon) in the atmosphere are probably equal to the amounts of these gases incorporated in the earth at its formation. Once they were released they remained in the atmosphere because they are geochemically inert and do not condense at earth temperatures.

The only other gas with an abundance greater than two parts per million is ozone (O_3). Its abundance varies with altitude, reaching a maximum of 12 parts per million at 30 kilometers. The amount of ozone in the atmosphere is set by the balance between the reactions by which it is created and those by which it is destroyed. The oxygen atoms formed by the photodissociation of mo-

DUST CLOUD INJECTED INTO THE STRATOSPHERE by the eruption of the volcano El Chichón in Mexico on April 4, 1982, was mapped by *Solar Mesosphere Explorer*, a polar-orbiting satellite operated by the University of Colorado at Boulder for the National Aeronautics and Space Administration. The sequence of computer-generated images on the opposite page shows the extent of the cloud at intervals of roughly a month beginning on April 2, immediately before the explosion. The color in the images corresponds to the amount of infrared emission from the cloud, changing from blue to green to red to yellow with increasing radiance. The cloud reached its greatest density in June (*third image*). Thereafter it spread slowly southward and began to disperse. Apparently the pattern of stratospheric circulation kept the dust from moving north of 30 degrees north latitude. The dust particles are highly reflective and decrease the amount of solar radiation reaching the surface of the earth. The net effect of the cloud will therefore be to decrease the mean global temperature (by as much as one degree Celsius). The temperature reductions associated with volcanic eruptions seem to be closely correlated with the sulfur content of volcanic debris. Data from the satellite *Nimbus 7* indicate that El Chichón injected unusually large amounts of sulfur dioxide into the stratosphere. The eruption is therefore expected to cause a relatively large drop in temperature for one of its size.





GLOBAL ENERGY BALANCE of the earth and the atmosphere determines the mean global surface temperature and the effective radiating temperature of the earth: the temperature it would appear to have to an observer in space. On the average the earth's surface transfers to the atmosphere an amount of energy equal to the amount of energy it absorbs; the value of the mean global surface temperature, which is about 13 degrees C., is that needed to keep the earth and the atmosphere in thermal equilibrium. On the average the earth as a whole emits into space an amount of radiant energy equal to the amount of radiant energy absorbed by the atmosphere and the earth's surface; the value of the earth's effective temperature, which is about −18 degrees C., is that needed to keep the earth in thermal equilibrium with space. The surface temperature is higher than the effective temperature largely because incident radiation is absorbed at altitudes lower than those from which radiation is emitted into space: the atmosphere is relatively transparent to radiation at visible wavelengths, where the solar emission spectrum peaks, and relatively opaque to radiation at infrared wavelengths, where the radiant emission from the earth is concentrated. (It should be noted that the earth emits about 114 units of infrared radiation; the net infrared emission is the difference between this value and the 93 units absorbed by the atmosphere and emitted back to the surface of the earth.)

lecular oxygen (O_2) by ultraviolet radiation react with molecular oxygen to form ozone. Ozone is destroyed by a number of reactions. Reactions in which "odd oxygen" (either O or O_3) is consumed and molecular oxygen is created lead to a net decrease in the concentration of ozone. Ozone is the only atmospheric gas that absorbs in the near ultraviolet (at wavelengths of from .2 to .3 micrometer) and therefore it plays a crucial role in shielding the surface of the earth from harsh radiation.

The composition of the atmosphere, together with the earth's distance from the sun, determines the atmosphere's energy budget, which in turn determines everything from the temperature of the earth's surface to the pattern of atmospheric circulation that redistributes solar energy over the surface.

The solar flux (the amount of energy carried by sunlight passing perpendicularly through a unit area) at the outer edge of the earth's atmosphere when the earth is at an average distance from the sun in its orbit is 1,367 watts per square meter. It has long been suspected that this "solar constant" is in fact a variable, but only in the past five years has it been possible to measure variations in the constant. Sensitive instruments aboard the *Nimbus 7* and *Solar Maximum Mission* spacecraft have shown that a large sunspot group causes a .1 percent drop in the received radiation. Larger variations may well be revealed by long-term measurements. It is estimated that a 1 percent change in the solar constant sustained for at least 10 years would raise the earth's mean surface temperature by about one or two degrees Celsius.

Not all the radiation reaching the earth is absorbed. About 30 percent of it is reflected back into space by the atmosphere and the earth's surface. The principal reflectors are clouds, atmospheric dust, the molecules of the atmospheric gases, snow and unvegetated land. The percentage of the incident radiation reflected back to space (the planet's albedo) could change substantially if the climate changed, if more dust were injected into the atmosphere by volcanic eruptions or if more land were cleared of forest. A decrease or increase in the albedo would lead to a net heating or cooling of the earth because the sunlight that is not reflected is absorbed.

The earth disposes of the solar energy it absorbs by emitting thermal radiation. Given the solar constant and the earth's present albedo the energy flux averaged over the globe must be about 240 watts per square meter if the earth is to remain in thermal equilibrium. The temperature of the average radiating level can be roughly estimated from the Stefan-Boltzmann law, which states that the flux of radiation given off by a black body (an ideal emitter) is proportional to the fourth power of its temperature in degrees Kelvin. That number (240 watts per square meter) is the power per unit area emitted by a black body at 255 degrees K. (−18 degrees C.). This is the average temperature of the atmosphere at an altitude of five kilometers.

A black body at 255 degrees K. emits radiation at a broad range of wavelengths with a flat maximum at about 12 micrometers (in the infrared region of the spectrum). The earth's surface and atmosphere behave approximately like black bodies and emit radiation in this range. Most of the radiation emitted by the surface is absorbed, mainly by water vapor, clouds, carbon dioxide, dust and ozone. The constituents of the atmosphere that absorb infrared radiation reemit it in all directions. Some of the emitted radiation is absorbed at the surface, adding to the heat energy there, some of it is reabsorbed by the atmosphere and some escapes into space.

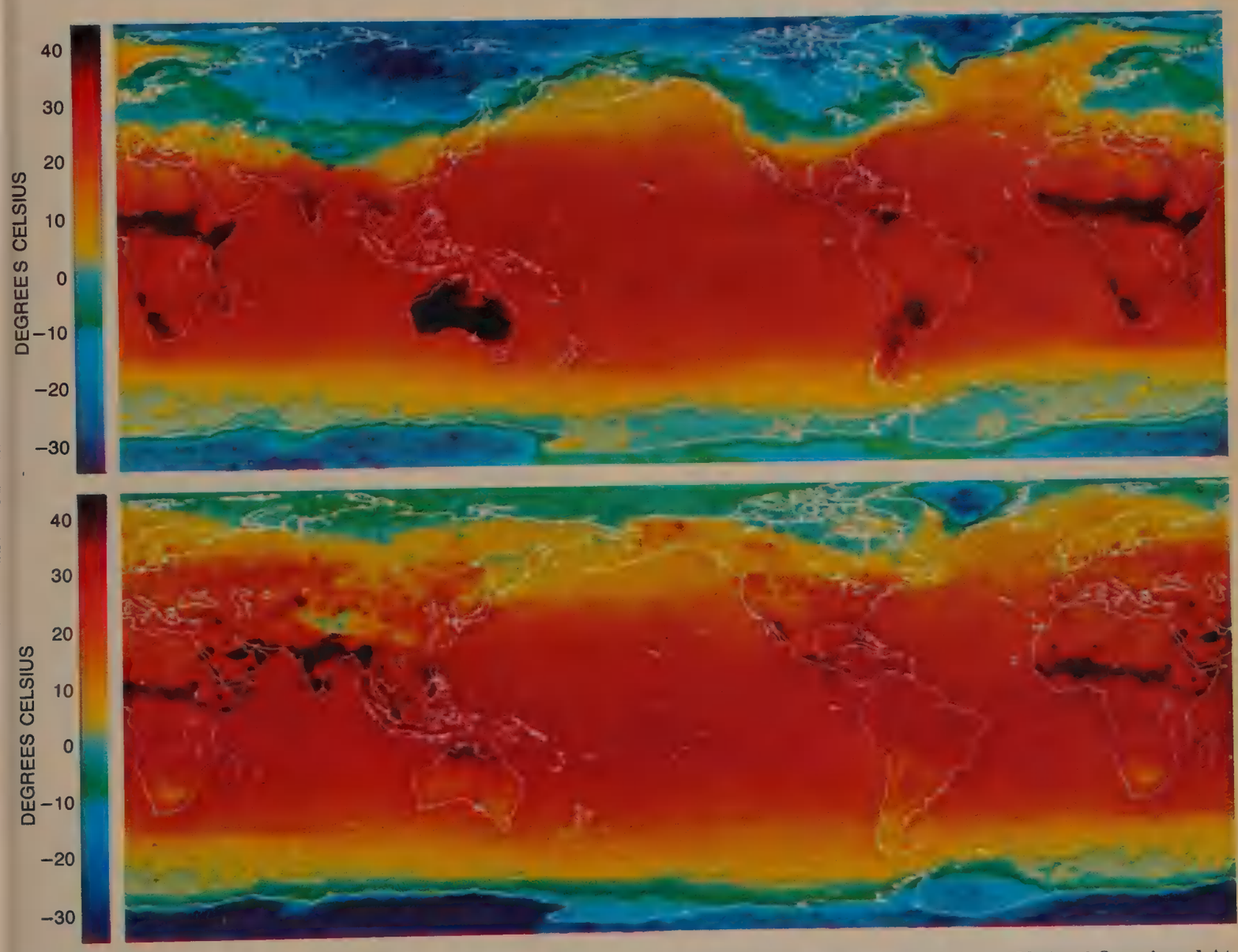
If the amount of an infrared absorber

such as carbon dioxide increases, more radiation is absorbed by the earth's surface and its temperature rises. In addition less of the radiation emitted by the surface escapes into space and the earth's stored thermal energy increases; thus infrared absorbers help to warm the earth. The role the atmosphere plays in heating the earth is traditionally called the greenhouse effect, but the term is somewhat misleading. Although the glass of a greenhouse does pass sunlight and block the escape of infrared radiation, most of the heating is due to the fact that the glass roof prevents heat from being dissipated by convection.

The mean global surface temperature (which is currently about 13 degrees C.) is the temperature needed to keep the earth's surface and the atmosphere in thermal equilibrium. The surface, heated by solar radiation and the infrared emitted downward by the atmosphere, transfers on the average an equivalent amount of energy to the atmosphere by evaporation, by conduction, by convection and by the emission of infrared. Local imbalances in the amount of energy absorbed and released by the surface, however, help to create vertical and horizontal temperature gradients in the atmosphere.

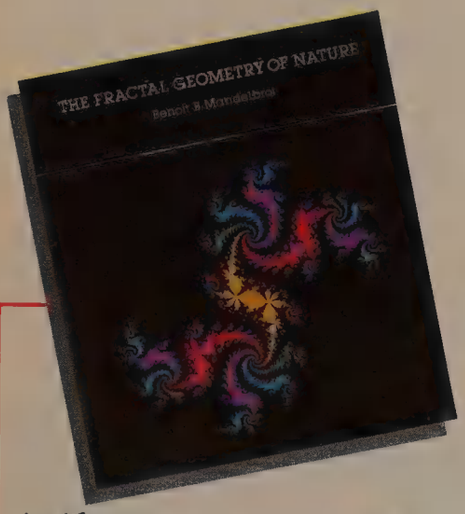
The earth absorbs more sunlight at low latitudes than it does at high latitudes. More than half of the sunlight is absorbed by the earth's surface; the rest is absorbed by the atmosphere. The atmosphere at low latitudes and altitudes therefore receives more energy from the earth's surface and from the sun than the atmosphere at high latitudes and altitudes. Moreover, the atmosphere loses more energy to space at high altitudes than it does at low altitudes. As a result of these imbalances temperatures in the atmosphere generally decrease from the Equator to the poles and from low altitudes to high altitudes. The temperature gradients drive the circulation of the atmosphere. The winds generally carry heat down the gradients, from relatively warm areas to relatively cool ones.

The temperature gradients and therefore the large-scale pattern of atmospheric circulation vary with latitude. Within 35 degrees latitude of the Equa-



TEMPERATURE GRADIENTS between the Equator and the poles and between the continents and the ocean drive the large-scale motions of the atmosphere. These maps of the mean surface temperature in January (*top*) and May (*bottom*) of 1979 illustrate the gradients, together with the change in them with season. For example, the formation of hot regions in northern India in May is a precursor of the southwesterly winds that blow from the cooler Arabian Sea across much of the Indian peninsula in the monsoon season from June to October. The increase in sea-surface temperatures from winter to summer is much less than the increase of land-surface temperatures largely because wave motion distributes heat to greater depths in the sea. These images were generated from data collected by the infrared and microwave sounding units carried by National Oceanic and Atmospheric Administration weather satellites. The surface temperature was derived from the microwave data by means of an analytic method that compares the data from the microwave sounder with those from the infrared sounder in order to distinguish among the contributions to the measured radiances made by the clouds, the atmosphere and the surface. The method was originally developed by Moustafa T. Chahine of the Jet Propulsion Laboratory of the California Institute of Technology for the purpose of analyzing atmospheric data acquired by space missions to other planets. It has since been applied to satellite data on the earth's atmosphere by Milton Halem and Joel Susskind of NASA's Goddard Space Flight Center.

Facts as enthralling as fiction



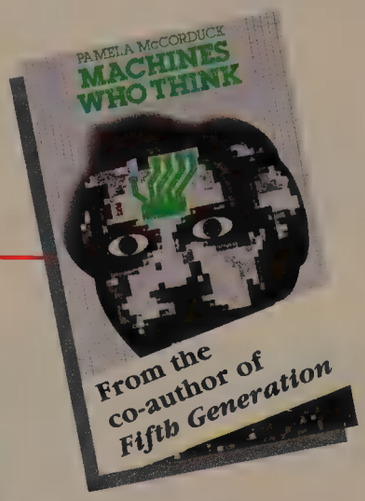
The Fractal Geometry of Nature

BENOIT B. MANDELBROT
IBM Fellow, Thomas J. Watson Research Center

Clouds are not spheres, mountains are not cones, and lightning does not travel in a straight path, but never before has geometry been able to represent the complexity of nature's shapes. In conceiving the new computer-generated geometry of fractal shapes, Benoit Mandelbrot succeeded in integrating nature, art, science, and geometry itself.

"[A] spectacular book loaded with beautiful illustrations from many fields of knowledge and graced by fascinating details about the lives of those who contributed to the subject. . . . This book is a 'must' for all who delight in surveying the frontiers of knowledge."
—John Archibald Wheeler,
American Journal of Physics

Clothbound, 460 pages, $32.50
ISBN 0-7167-1186-9



Machines Who Think

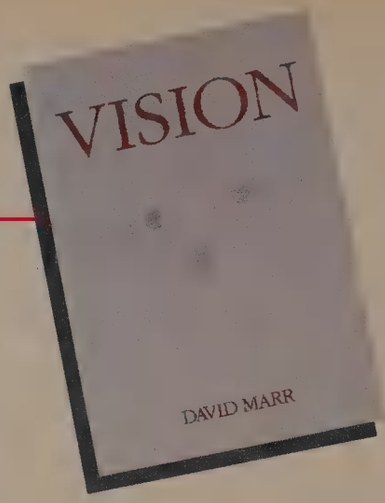
PAMELA McCORDUCK
University of Pittsburgh

Literate and well informed, MACHINES WHO THINK traces the course of artificial intelligence from its mythic beginnings to its sometimes frightening future. Told in the words of those who developed the field, MACHINES WHO THINK describes the evolution of artificial intelligence from the dreams of the past to the reality of the present —and beyond.

"Without doubt the best book on the subject so far published for the general reader."
—*The Washington Post*

"The book's magnetism is McCorduck's facility with language and structure: a storytelling tradition that seduces us into a love affair with a most important story."
—*Creative Computing*

Paperbound, 375 pages, $10.95
ISBN 0-7167-1135-4



Vision

DAVID MARR
Late of the Massachusetts Institute of Technology

Available for the first time in paperback, VISION presents an entirely new way of looking at sight. Based on the computational theory of vision the author developed at MIT, the book focuses on the information-processing tasks the eye performs and relates them to the fields of computer science and neuroscience.

"A milestone in the history of the subject. Even if no single one of Marr's detailed hypotheses ultimately survives, which is unlikely, the questions he raises can no longer be ignored. Marr's lifework will have been vindicated when neuroscientists cannot understand how it was ever possible to doubt the validity of his theoretical maxims."
—H. C. Longuet-Higgins,
Science

"I rate the book as brilliant. Reading it created in me the same unique sense of excitement and beauty I felt on first learning about the early experiments of Hubel and Wiesel."—Denis Baylor, Stanford University Medical Center

Paperbound, 397 pages, $19.95
ISBN 0-7167-1567-8

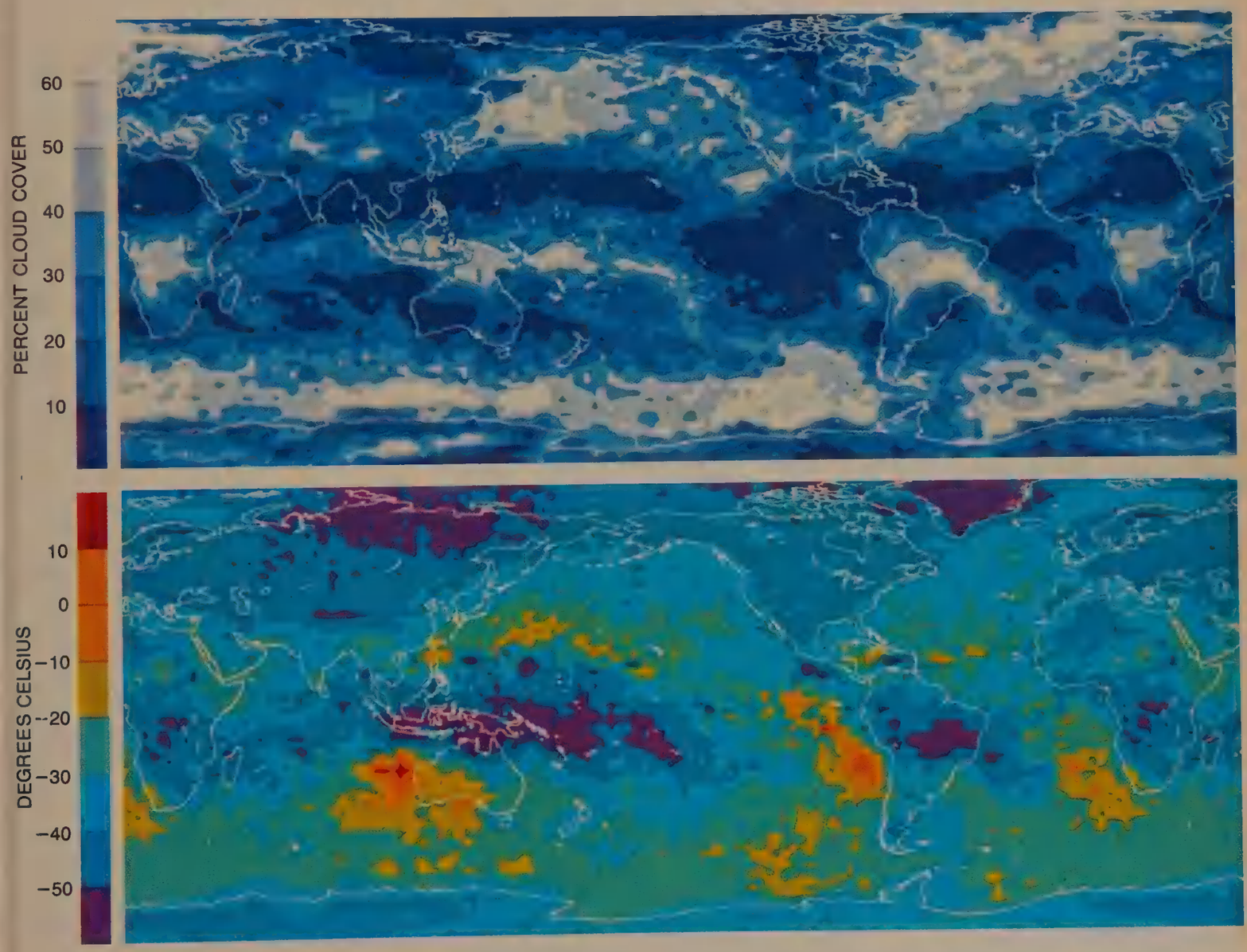
To order your copies, simply fill in the attached card or send payment (including $1.50 for shipping and handling; California, New York, and Utah residents add appropriate sales tax) to: W. H. FREEMAN AND COMPANY, 4419 West 1980 South, Salt Lake City, Utah 84104.

W. H. FREEMAN AND COMPANY

tor the large-scale winds are meridional (north to south). Two huge circular cells connect the hot moist air rising over the Equator to the warm dry air sinking at higher latitudes. The hot air rising in what is known as the Intertropical Convergence Zone (ITCZ), within 10 degrees north and south of the Equator, cools as it rises, which results in intense precipitation. The ITCZ coincides with the tropical rain forests and regions of heavy precipitation over the oceans. The dry air spreads poleward at high altitudes and sinks throughout the subtropics, the regions between 10 and 35 degrees latitude in both hemispheres. When it reaches the surface again, it is still dry; the great deserts of the world are in these zones.

Poleward of 35 degrees latitude the circulation is predominantly zonal rather than meridional. The flow of air is dominated by the meandering high-altitude jet stream that blows generally eastward in both hemispheres at speeds of up to 45 meters per second (100 miles per hour). Long waves or undulations that begin as small fluctuations and deepen with time are superimposed on the general west-to-east flow. The troughs of the waves are where the jet stream is closest to the Equator; the ridges are where it is closest to the poles. As the waves develop, masses of cool air move toward the Equator behind (to the west of) the troughs and masses of warm air move toward the poles ahead of (to the east of) the troughs. The development of the waves thus contributes to the transport of warm air toward the poles and of cold air toward the Equator. The masses of air trapped in the troughs and ridges are carried eastward by the movement of the high-altitude waves. The waves thereby help to determine the movement of the surface high- and low-pressure regions that dominate the weather over North America and other mid-latitude regions.

Beginning about 50 years ago the dynamics of large-scale atmospheric motions (as they are now understood) began to be formalized in mathematical equations. Among the pioneers in this work were Carl-Gustav Rossby, the Swedish meteorologist for whom long waves in the jet stream are named, and



MAP OF MEAN CLOUD COVER in January, 1979 (*top*), generated by the image-processing group at the Jet Propulsion Laboratory from data collected by weather satellites provides an index to characteristic atmospheric circulation patterns and the climate zones they create. For example, the belt of clouds near the Equator is created by the convergence of the updrafts of two huge north-south circulation cells called Hadley cells. The hot air rising in this zone, called the Intertropical Convergence Zone (ITCZ), cools as it rises. As the map of the mean temperature at the top of the clouds in January, 1979 (*bottom*), shows, the cloud tops over the ITCZ are high and therefore cold. The tropical rain forests and comparable regions of heavy precipitation over the oceans are in this zone. The annual precipitation can reach two meters and exceeds the annual evaporation by more than 50 centimeters. The relatively cloud-free strips above and below the ITCZ extending up to about 35 degrees latitude are the regions over which the still-warm but by now dry air in the Hadley cells sinks. The great deserts of the world are in these zones, which are called the subtropical arid zones. The large-scale circulation pattern at higher latitudes is governed by the development of waves in the meandering high-altitude jet streams and is therefore more variable than the pattern at low latitudes. The pattern of waves in the zonal flow at middle latitudes is suggested by such features of the maps as the deflection of the cloud belts (storm tracks) southward along the west coast of North America and northward along the east coast.

Jule G. Charney, who was for many years at the Massachusetts Institute of Technology. Rossby and Charney led the way to the development of workable computer models of the large-scale circulation by developing equations that filtered out the effect of small-scale motions. Earlier atmospheric models had been subject to the explosive growth of errors in part because no such discrimination was made. Today global atmospheric models running on high-speed computers provide the basis for regional and local daily weather forecasts. The limit on useful forecasts is now about a few days to a week and is proving difficult to extend.

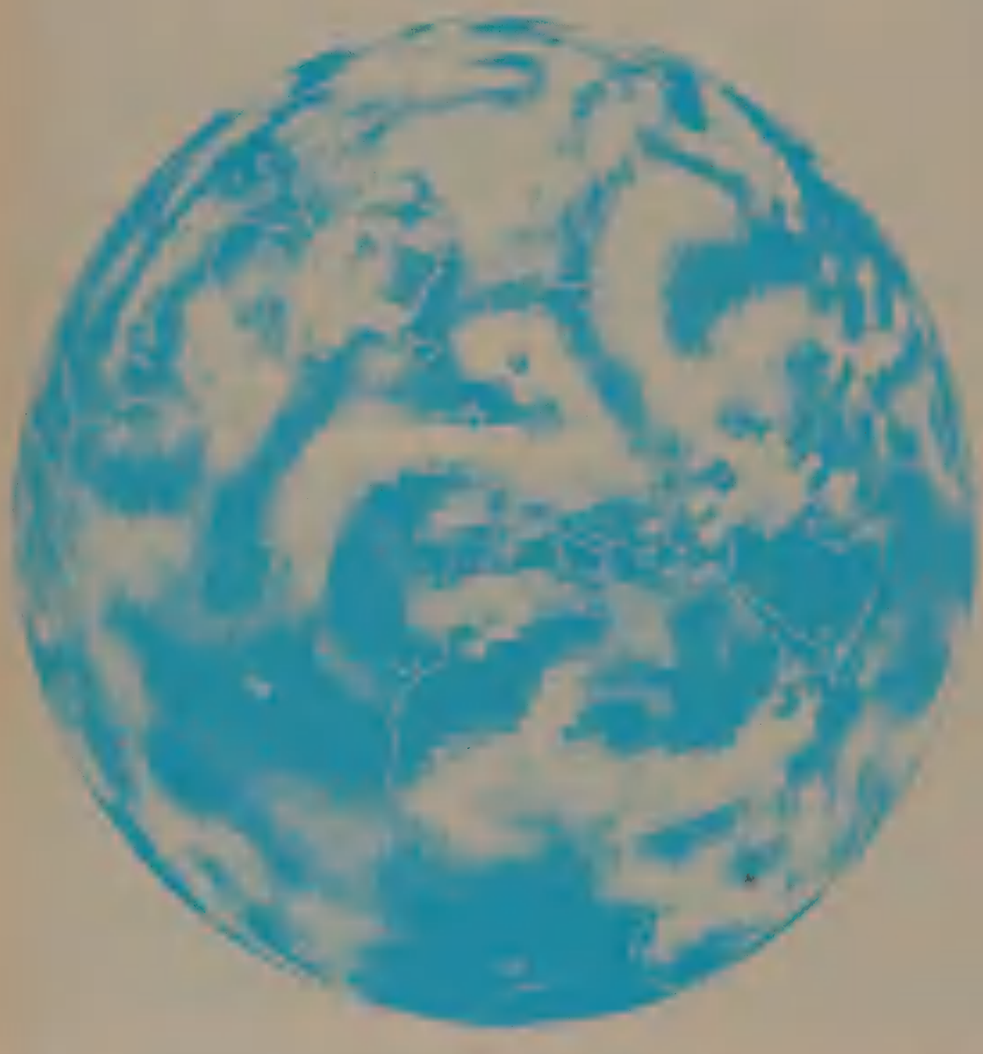
The first step in developing a forecast is to summarize the values of different atmospheric variables, such as wind speed and pressure measured at a large number of scattered locations, as the average values that apply at the intersections of an imaginary three-dimensional grid wrapped around the globe. The mathematical model, which is based on simple physical principles such as Newton's second law of motion and the gas law, is then used to determine how conditions at each point on the grid will change in a short interval of time, such as 10 minutes, during which the rate of change can be assumed to be constant. The new values are then used in place of the original ones and the process is repeated through the forecast period.

One of the major problems weather forecasters face, the problem presented by small-scale atmospheric phenomena, may suggest why it is difficult to extend the period of useful forecasts. Small-scale phenomena such as turbulent eddies and convective storms are relatively energetic, as Herbert Riehl of the National Center for Atmospheric Research and Joanne Starr Malkus of the Woods Hole Oceanographic Institution pointed out in the 1950's. For instance, most of the vertical motion in the Tropics occurs in isolated storms covering .1 percent of the area. The latent heat of evaporation released when water condenses in the storm clouds is a major source of atmospheric energy. In addition the clouds reflect sunlight and absorb infrared radiation, thus further altering the distribution of energy that eventually determines the weather. Collectively these small-scale phenomena have a significant effect on global weather patterns, although individually they are below the level of resolution of the weather-forecasting models.

The small-scale phenomena are not ignored in the models; they are simply inadequately treated. The equations that are applied to tracking large-scale motions in the atmosphere are deterministic; the values of the parameters in the grid at any one time determine those in the grid after the next cycle of calculations. Small-scale processes, however, are handled by statistical subroutines of the models that specify the most probable net effect of all small-scale processes inside a cube on the basis of the values at the cube's corners. Sometimes the most probable events do not occur, and statistical errors therefore contribute significantly to the rate at which the evolution of the model diverges from that of the atmosphere.

It is unlikely that this problem will be solved simply by increasing the level of resolution of the atmospheric models. The resolution of a model, which is determined essentially by the size of a cube in the grid, is ultimately limited by the number of arithmetic operations needed to keep track of the evolution of the variables in the grid. For example, keeping track of seven atmospheric variables (temperature, pressure, water vapor, cloud cover and the wind speed along three axes) in a grid consisting of cubes 200 kilometers square and 10 layers deep actually means keeping track of a million variables. Interactions with nearby variables largely determine how each variable changes with time. Some 500 arithmetic operations are required to compute the interactions to which one variable is subject, so that a total of about .5 billion operations must be done for each 10-minute time step of the model. A tenfold increase in the spatial resolution of the three-dimensional grid would increase the number of variables by three factors of 10, and the number of time steps per hour would have to increase by a comparable factor. The number of arithmetic operations would therefore increase 10,000-fold, and yet such a model would still not resolve atmospheric phenomena smaller than 20 kilometers.

The problem of small-scale atmospheric phenomena and other problems that limit the period of useful forecasts



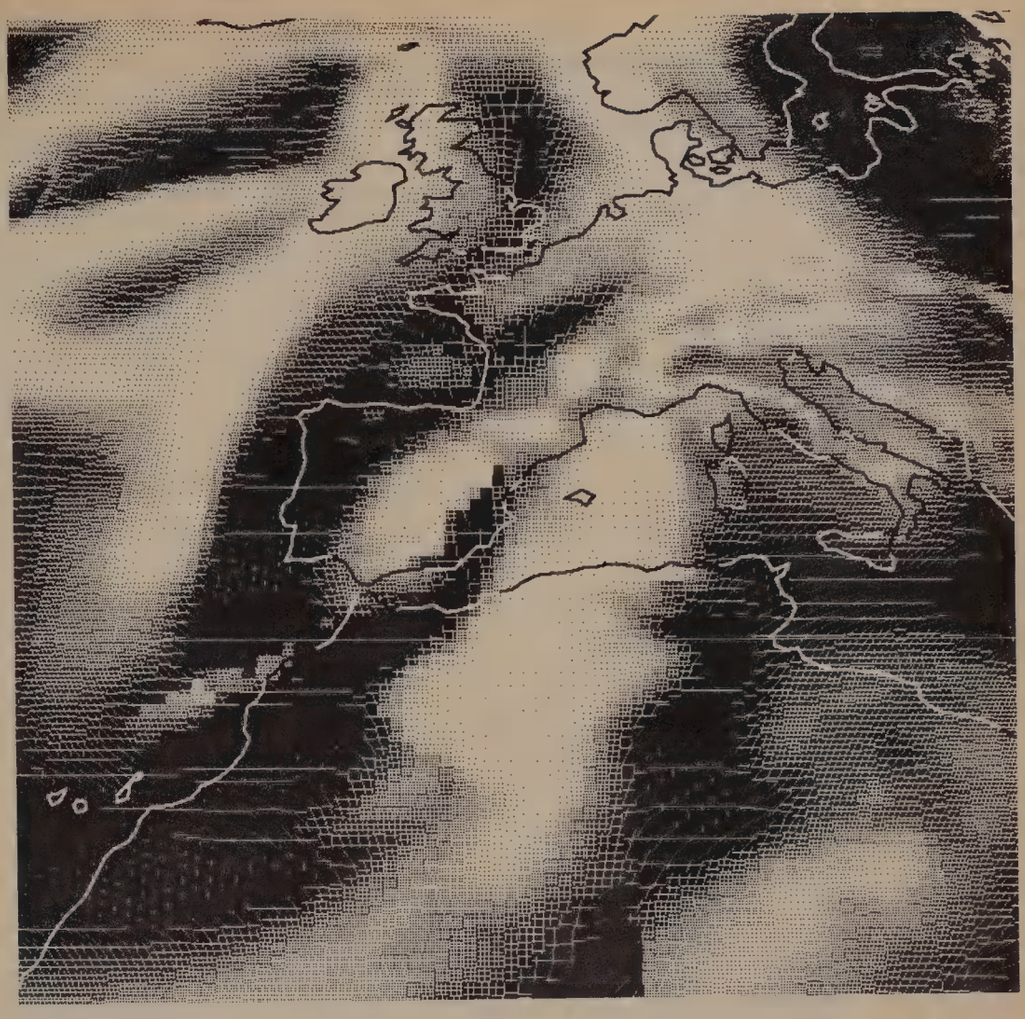
GLOBAL ATMOSPHERIC MODEL developed by the European Centre for Medium Range Weather Forecasts, an organization supported by 17 European nations, produced the data from which this image of predicted global cloud cover was generated. Measurements of atmospheric variables such as temperature, pressure, wind speed and moisture content, made at many locations scattered around the world, are summarized as the average values that would hold at the intersections of a three-dimensional grid wrapped around the globe. The model calculates future conditions from the initial conditions by means of equations ultimately based on simple physical principles such as Newton's laws of motion, the thermodynamic equations and the gas law. The resolution of the image suggests the limit to the resolution of global forecasting models imposed by the enormous number of calculations needed to keep track of the evolution of the data points. For example, the European Centre's Cray-1 computer must do 500 billion arithmetic operations to predict the behavior of the atmosphere 10 days later.

are being addressed by a number of major research programs. The experiments being conducted as part of the Global Atmospheric Research Program (GARP), organized by the World Meteorological Organization and the International Council of Scientific Unions in the early 1960's, are basically designed to provide data about atmospheric phenomena that will allow better forecasting models to be developed. The GARP Global Weather Experiment conducted in 1979, for example, consisted of two two-week intensive observation periods. During these periods atmospheric variables were measured by balloons, satellites, aircraft, ships and buoys. Parameters of the sea, land, ice, soil and vegetation that might have an impact on the weather were measured as well.

One goal of the experiment was to see how sensitive the forecasts for one part of the globe, say North America, are to initial conditions over other parts of the globe, say the normally ill-defined conditions over the South Pacific. Another was to determine whether measurements of variables, such as soil moisture, that are not now exploited in weather forecasting would improve forecasting accuracy. A third goal was to provide more detailed data on the troublesome small-scale atmospheric phenomena.

Among the research programs in the U.S. directed at improving global forecasting models are those at the National Center for Atmospheric Research in Boulder, Colo., the Goddard Space Flight Center of the National Aeronautics and Space Administration and the Geophysical Fluid Dynamics Laboratory of Princeton University. One of the problems workers at these institutions are addresssing is that of determining which of a multitude of possible sources of error is actually responsible for erroneous forecasts.

Is it the inaccuracy of the initial observations, the limitations of the models or the unpredictability of the atmosphere itself that is currently setting the upper limit on the period of useful forecasts? There is no simple answer to the question, but a test applied by Edward N. Lorenz of the Massachusetts Institute of Technology, who was one of the first to make estimates of this kind, may suggest what can be done to pin down the sources of inaccuracy. If computations are run on two sets of initial weather conditions that differ by an amount within the estimated uncertainty of the observations, the rate at which the calculated conditions then diverge yields a measure of the forecasting inaccuracy due primarily to observational error. Current atmospheric models double small differences in initial conditions every two to three days. Assuming that the observational errors are distributed uniform-



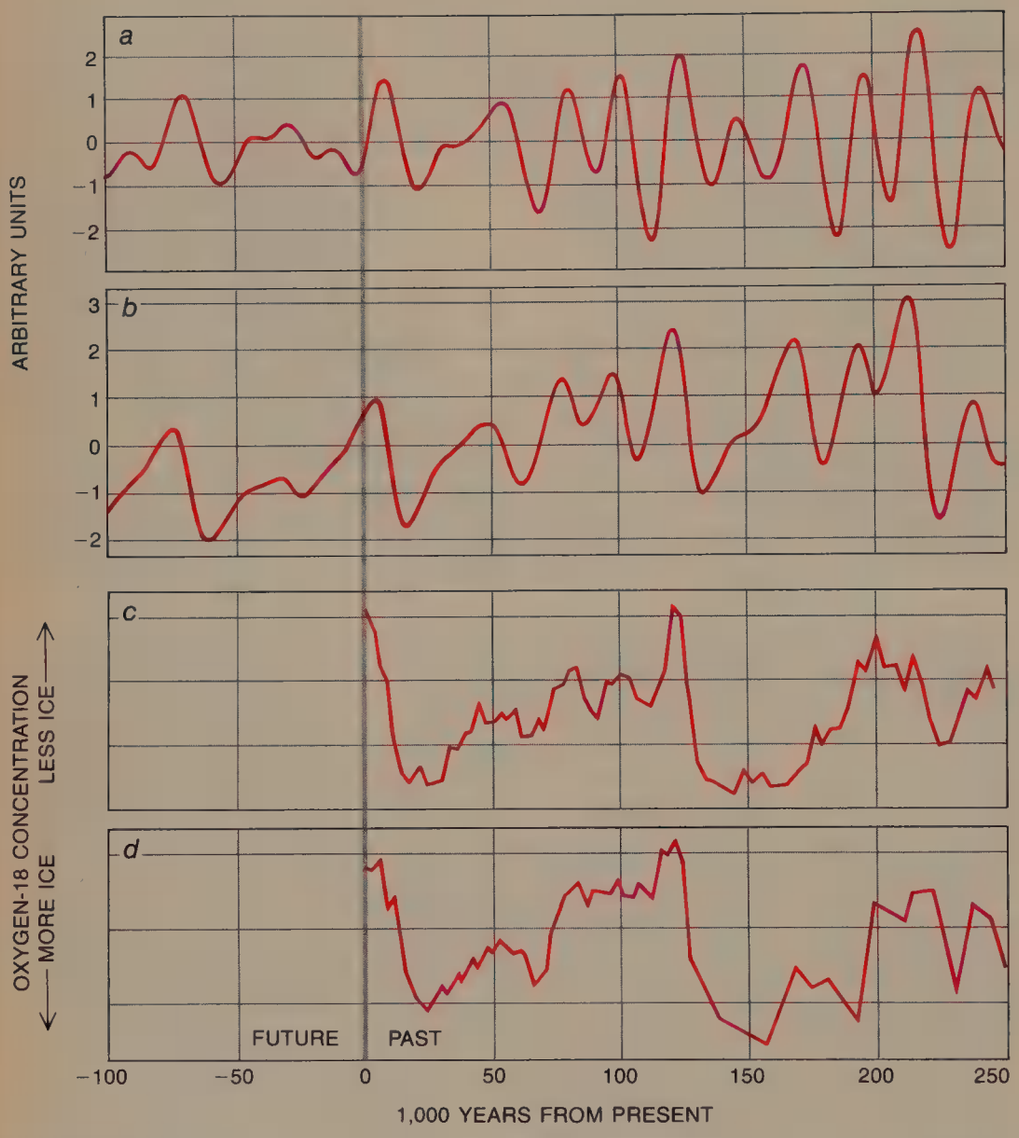


HIGH-RESOLUTION ONE-DAY FORECAST of cloud cover produced by a limited-area version of the European Centre for Medium Range Weather Forecasts' atmospheric model (*top*) compares well with an image of cloud cover made by the satellite *Metosat* one day later (*bottom*). The grid used for limited-area forecasts consists of cubes 50 kilometers square (less than .5 degree in latitude or longitude) and is 15 layers deep. The cloud system approaching Europe from the Atlantic is a cold front: advancing cold air is underriding retreating warm air.

ly over the globe and that the evolution of models exactly duplicates the evolution of the weather, this doubling rate suggests one-to-two-week forecasts (several doubling times) are theoretically possible.

How can forecasting models be improved? Given the complexity of the system the models seek to simulate, the best approach may be not more exhaustive calculations but rather calculations discriminating more accurately between phenomena that have a significant impact on the development of the weather and phenomena that have an insignificant impact, in effect improved recognition of patterns. Patterns that persist for more than a few days, called blocking phenomena, might include waves in the jet stream, the paths that storms habitually take or areas of persistent drought.

Conditions in the upper levels of the ocean may also provide a basis for long-range weather forecasting. Anomalies in sea-surface temperatures and in the slope of the sea caused by the prevailing winds can persist for months or years. Anomalies in the surface temperature could be exploited to gauge the amount of energy that will be supplied to the atmosphere as the latent heat of evaporation. More generally, such anomalies can trigger chains of events that unfold in a predictable way, such as the sudden change in ocean currents and weather patterns near the coastline of Peru called an El Niño event. The major El Niño event of 1983 is providing an opportunity to study the causes and possible consequences for global weather patterns of such changes in ocean temperature and circulation patterns.



MILANKOVITCH THEORY of the climate cycles during the recent ice age holds that the forcing mechanism is the change in the amount of Northern Hemisphere summertime insolation (incident solar radiation) caused by the variation of three parameters of the earth's orbit around the sun. The two parameters that determine the period of fluctuations in Northern Hemisphere insolation are the tilt angle of the axis, or the angle between the axis and the plane of the ecliptic (which determines to what extent the North Pole is pointed toward the sun in the Northern Hemisphere summer), and the direction in which the earth's axis points (which determines whether perihelion, the time of the year when the earth is closest to the sun, comes in the Northern Hemisphere summer or the Southern Hemisphere summer). The orbital input (*a*) fluctuates with these parameters, which have periods of 40,000 and 20,000 years. The ratio of the heavy isotope oxygen 18 to the lighter isotope oxygen 16 in ocean sediments provides a measure of the global ice volume. As saturated air moves poleward the water molecules incorporating the heavier isotope are preferentially removed by precipitation, so that snow falling at high latitudes is enriched in oxygen 16. The oceans are therefore enriched in oxygen 18 in cold periods when snow accumulates on the land. The isotope ratio in two representative deep-sea cores, one from the southern Indian Ocean (*c*) and the other from the Pacific (*d*), measured by John Imbrie and his son John Z. Imbrie of Brown University, does fluctuate with periods of 40,000 and 20,000 years but would be better characterized as a sawtooth curve with a period of about 100,000 years. A climate model developed by the Imbries (*b*), which assumes a long lag time before ice accumulates in response to the decrease in insolation and a much shorter lag time before the ice melts in response to an increase in insolation, suggests how the changes in incident solar radiation might be translated into the changes in global ice volume.

Although the weather is highly variable, large deviations in temperature and precipitation are usually followed by a return to average or normal values. In other words, the long-term weather or climate seems relatively stable. The geologic evidence, however, indicates that the earth's climate does change radically on the scale of thousands and millions of years.

Perhaps the most dramatic example of such a change is the recent ice ages. For most of its history the earth seems to have been largely ice-free. Beginning several million years ago, however, it started to undergo cyclic glaciations. Over periods ranging from 20,000 to 100,000 years ice sheets build up on the land masses of the Northern Hemisphere. (Southern Hemisphere glaciers grow at the same time, but they account for much less of the global ice volume because there are no large land masses from 40 to 70 degrees south latitude on which glaciers can grow.) In the Northern Hemisphere the ice builds up to a thickness of three kilometers, and the weight of the ice depresses the crust under it by as much as 500 meters. When the ice becomes thick enough, the pressure of its weight causes it to flow outward and southward until it covers large areas of North America and Eurasia. The sequestering of water in the ice lowers the level of the oceans by 100 meters or more. At the height of a glacial period the average temperature of the earth is two or three degrees C. lower than normal. The ice then retreats abruptly. The length of the retreat phase is uncertain, but the geologic evidence suggests it is several thousand years or less.

What causes the climate to go on such a rampage? The answer seems to lie in the response of the atmosphere to external forcing mechanisms such as changes in the amount or distribution of solar radiation. The external changes by themselves are too small to account for the large changes in climate; what appears to matter more are how these

changes are amplified or damped by the atmosphere, the ocean and the ice. Among the first to attempt to assess the effects that might increase or decrease the sensitivity of climate to external forcing were Mikhail I. Budyko of the Main Geophysical Observatory in Leningrad and William D. Sellers of the University of Arizona. In the 1960's they independently published climate models incorporating feedback effects that governed the climate changes triggered by changes in the solar constant. For example, lower polar temperatures would cause snow and ice to accumulate at high and middle latitudes. The albedo of snow is higher than that of land or water, and so less sunlight would be absorbed at these latitudes, causing the temperatures to drop further.

Formulating models of climate variability remains an uncertain business. For one thing the relative importance of positive and negative feedback effects is difficult to assess. For example, ice-albedo feedback and other factors accentuating the decrease in temperature at high latitudes might be offset by atmospheric and oceanic circulation driven by the gradient of temperature between the Equator and the poles. Climate models are also subject to many of the problems encountered in weather models, problems exacerbated by the relative lack of data on conditions in glacial epochs. In addition climate models must take account of the interaction of the atmosphere with the ocean and the ice sheets, interactions that can be ignored with relative impunity in short-term weather models. Little is known, however, about the dynamics of the deep ocean and the continental ice sheets.

In spite of these uncertainties, comparing climate models with the geologic record of climate change is instructive. The most successful models of the recent ice ages are based on the astronomical theory of climate forcing. This theory concerns changes in the amount of solar energy received by the earth that are associated with three cyclically changing parameters of the earth's orbit around the sun. The first parameter is the tilt of the earth's axis (the angle between the axis and the plane of the orbit), which has a period of about 40,000 years. The second is the direction in which the earth's axis points, which has a period of about 20,000 years. The third is the eccentricity of the earth's orbit (the degree to which it departs from a circle), which has a period of 100,000 years. The orbital forcing can be calculated for millions of years into both the past and the future.

The astronomical theory is also known as the Milankovitch theory, after Milutin Milankovitch, the Yugoslav scientist who made such calculations in the 1920's and 1930's. According to Milankovitch, the key factor for the earth's climate is not the total amount of sunlight received by the globe in the course of the year, which in any case changes only slightly with changes in the orbital parameters, but rather the amount of sunlight received at high latitudes in the Northern Hemisphere during the summer. The tilt angle, which determines to what extent the North Pole is pointed toward the sun during the Northern Hemisphere summer, has the greatest impact on the amount of sunlight received in the summertime. The direction in which the axis points determines whether perihelion, the time of year when the earth is closest to the sun, comes during the Northern Hemisphere summer or the Southern Hemisphere summer. The eccentricity of the orbit affects the amount of sunlight received at perihelion and thus determines the amplitude of the 20,000-year forcing cycle. The eccentricity of the orbit by itself has very little effect on summertime insolation and so there is little climate forcing at the period of 100,000 years.

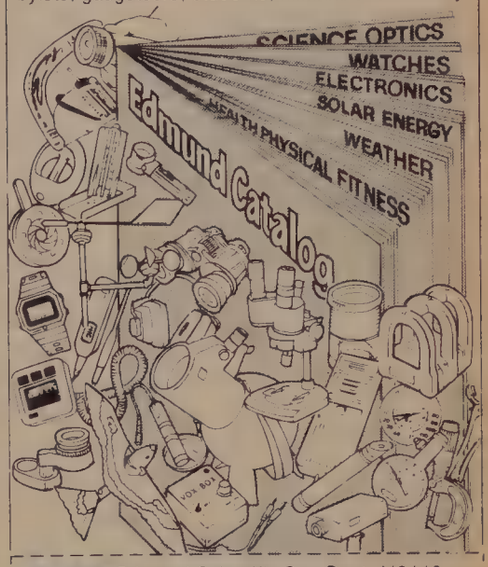
The clearest corroborating evidence for the orbital forcing of the glacial cycles comes from the ratio in ocean sediments of two isotopes of oxygen: oxygen 18 and oxygen 16. As an air mass moves away from the Equator water molecules incorporating oxygen 18 are preferentially removed by precipitation. The snow that falls at high latitudes holds more of the lighter isotope. Therefore when snow accumulates on the land, the oceans are slightly enriched in oxygen 18. Changes in the oxygen-isotope ratios in deep-sea sediments thus reflect changes in global ice volume. One of the first to recognize the importance of the sedimentary isotope ratios as an index to past changes in the climate was Cesare Emiliani of the University of Miami. The results of systematic oxygen-isotope analyses were published in 1976 by J. D. Hays of Columbia University, John Imbrie of Brown University and Nicholas J. Shackleton of the University of Cambridge.

The isotope record of the past 500,000 years shows prominent oscillations in the global ice volume at periods of 40,000 and 20,000 years. The dominant feature, however, is a steady increase in the ice volume for 100,000 years or so followed by a sudden decrease. Climate models in which the ice simply responds passively to the orbital forcing do not duplicate the isotope record; the ice volume predicted by the models oscillates with periods of 40,000 and 20,000 years in direct response to the forcing. Some other factor or factors must be introduced to account for both the 100,000-year cycle and the abrupt recovery from glaciation.

Models recently published by David

Send for Your Free Edmund Scientific Catalog Today!

America's largest mail-order market place of thousands of unusual and fascinating values for hobbyists, gadgeteers, students, schools and industry.



Edmund Scientific Co., Dept. N2110
Edscorp Bldg., Barrington, NJ 08007
Please send me your FREE Catalog.

NAME

ADDRESS

CITY STATE ZIP

DRESS FOR SUCCESS... FOR LESS.



Send for free catalog of fine traditional clothing at significant savings.

Shop conveniently at home or office, from our 100-page catalog for men and women. It's brimming with classic clothing and accessories, shown in full color. Everything is priced far below most fine stores, and your satisfaction is unconditionally guaranteed.

J.S. A. Bank® Clothiers

109 Market Place/Baltimore, MD. 21202
(301) 837-8838

Please send free catalog to:

Name _____

Address _____

City _____

State _____ Zip _____

© Copyright 1983 Jos. A. Bank Clothiers, Inc. BEJ

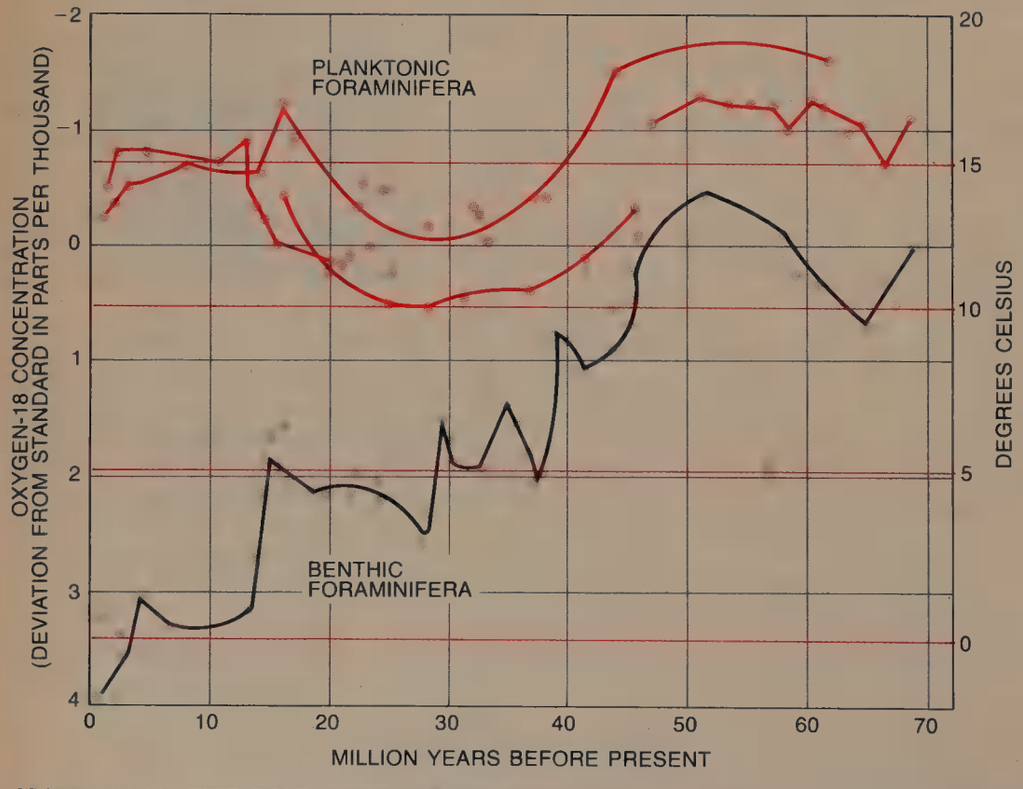
Pollard of Oregon State University illustrate the problem and a possible solution. He finds that the earth's climate at present is such that the "normal" state of the earth is icebound. There should be extensive permanent ice sheets. In his "passive" model the global ice volume fluctuates slightly with periods of 40,000 and 20,000 years, but it does not undergo the dramatic losses at periods of 100,000 years recorded in the isotope record. In order to get a more realistic response Pollard introduced a mechanism of rapid ice loss into the model: the calving of icebergs into the ocean causes the ice sheet to become unstable when it reaches a critical size.

The earth's crust does not respond immediately to glacial loading. During extended periods of glaciation, however, the weight of the ice depresses the crust until it is below sea level, although the top of the ice is still much above sea level. At this point the ice is particularly vulnerable to small changes in incident sunlight, such as those that might be caused by orbital forcing. If the ice sheet thins in response to such changes, the crust does not rebound immediately. Along the outer edge of the ice 90 percent of its thickness may be below sea level. Since ice is lighter than water, if the sea intrudes into the crustal depression, the ice will float. Floating ice is assumed to suffer rapid loss along its outer edge as icebergs calve off the ice sheet into the shallow sea. The ice sheet thins further as the thicker ice at the center of the sheet flows outward to establish a new equilibrium. Rapid loss and the consequent retreat of the ice continue until the edge of the ice sheet reaches high ground.

There is evidence that the crust in areas such as Scandinavia and Canada from which the last glacier retreated some 10,000 years ago is still rebounding. This puts the crustal response time into the same range as the one assumed in the model. Still, whether the instability of the ice sheet brought on by the depression of the crust is what causes the retreat of the ice during interglacial epochs remains the subject of speculation.

What is known about the climate of the far distant past, before the ice ages? The climate models devised by Budyko and Sellers predict that the earth would turn into a bright, ice-covered planet if the solar output were a few percent lower. The feedback effects the models incorporate not only increase the earth's sensitivity to the decrease in solar output but also, once the earth cools off, decrease its sensitivity to subsequent increases in solar output. In short, if the solar constant were lower, the earth would be locked in a deep freeze. Has anything of the kind ever happened?

Astronomers know something about the sun's past because they can observe stars that are much younger than the sun but have the same mass and composition. They conclude that the solar output was 40 percent lower shortly after the earth formed some 4.6 billion years ago than it is today and that it has been increasing ever since. Theoretical models of the sun's internal structure and evolution support this conclusion.



CLIMATE OF THE DISTANT PAST was characterized by higher mean global temperatures than those recorded today. One paleothermometer is the ratio of oxygen 18 to oxygen 16 in the sediments formed by the shells of the microscopic marine animals called foraminifera. Measurements of isotope ratios made by Samuel M. Savin of Case Western Reserve University show the cooling trend at high latitudes during late Cretaceous and Cenozoic times. At the end of the Cretaceous the temperature of the bottom of the ocean, which is now about zero degrees C., was about 12 degrees. Temperatures during much earlier periods were even higher.

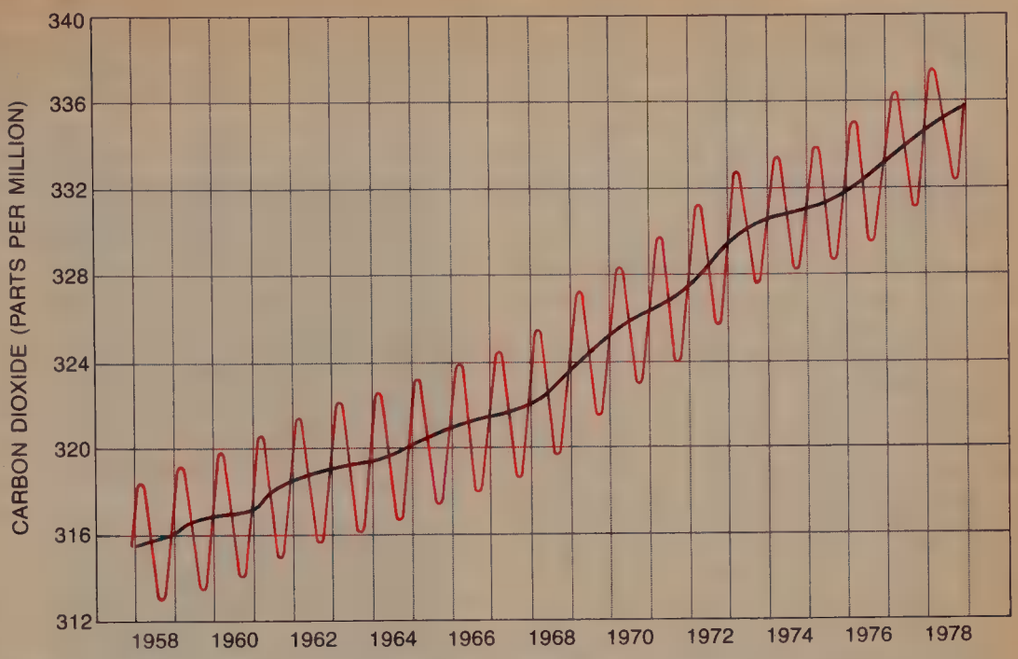
Geologists have compiled only a fragmentary record of the climates in the distant past, but the general trend of the evidence is clearly opposite to the one predicted by the trend of the solar constant and models of the earth's response to the forcing mechanism. Samuel Epstein and his colleagues at the California Institute of Technology have inferred temperatures from the oxygen-isotope ratios and comparable ratios of hydrogen 2 (deuterium) to hydrogen 1 (ordinary hydrogen) in cherts (crystalline siliceous sedimentary rocks). From the isotope ratios in a chert 3.5 billion years old they inferred that the temperature of the ocean in which the rock formed was higher than 50 degrees C. Different isotope ratios in more recent cherts suggest that temperatures became progressively lower. From the oxygen-isotope ratios in the sediments formed by the calcium carbonate shells of benthic foraminifera (minute animals of the deep ocean) it was inferred that the temperature at the bottom of the ocean was about 15 degrees C. in the period from 150 to 50 million years ago. Today temperatures at the bottom of the ocean are close to zero degrees.

Fossils from the past 600 million years also suggest that the earth has usually been warmer than it is today. Coal beds in Greenland and Antarctica indicate that tropical plants once grew there. The significance of such evidence is complicated by the drifting of the continents, which may have placed a given land mass in a latitude warmer than the one it is in today. Less ambiguous testimony is provided, however, by the fact that in some geologic periods, such as the Carboniferous of between 345 and 280 million years ago, coal beds were laid down all over the world.

There is no universally accepted explanation for the warmer temperatures of the past. The most likely causes, however, are continental drift and changes in atmospheric composition. Changes in the distribution of the continents could have a substantial effect on the climate because they would change the patterns of atmospheric and oceanic circulation. For example, 100 million years ago the Arctic Ocean was part of the Pacific. Winds and the ocean currents they drove might thus have kept the Arctic free of ice. As the Bering Strait closed, the freezing of the Arctic Ocean might have launched the earth into its present phase of cool climates.

Another possibility is that the atmosphere may have held far more carbon dioxide than it holds at present. Carbon dioxide is the most likely gas for this

role because large quantities of it are stored in surface reservoirs from which it might conceivably have been released into the atmosphere. If the atmosphere held 100 to 1,000 times more carbon dioxide than it holds at present, the increased reradiation due to the carbon dioxide, together with that due to the larger amount of water vapor in a warmer atmosphere, would have maintained the surface at temperatures 20 to 30 degrees C. higher than they are today. Since the carbon dioxide in the atmosphere is currently in equilibrium with the carbon compounds dissolved in the ocean and stored in ocean sediments, however, this hypothesis implies that there were major differences between the ocean's composition 100 million years ago and its composition today. Not enough is known about the composition of either the ocean or the atmosphere of the distant past to prove or disprove this hypothesis.



CONCENTRATION OF CARBON DIOXIDE in the atmosphere at a site near the summit of Mauna Loa on the Island of Hawaii has been measured continuously since 1958 by Charles D. Keeling of the Scripps Institution of Oceanography and the National Oceanic and Atmospheric Administration. The site was chosen because there was little local contamination of the air and because the air over this part of the globe is probably well mixed. Superimposed on the steady increase in carbon dioxide (*black*) are seasonal fluctuations (*color*) due to the uptake of carbon dioxide by Northern Hemisphere plants in the summer and the oxidation of plant tissues in the winter. The fluctuations are dominated by Northern Hemisphere plants because there is less land in the Southern Hemisphere and less seasonal change in plant life at the Equator. Measurements taken near the South Pole show a comparable increase in carbon dioxide.

The forcing mechanisms described so far are slowly varying and stimulate relatively slow responses. It is becoming clear, however, that not all changes in the climate are gradual. When a large meteoritic body strikes the earth or a large volcano erupts, significant amounts of dust and condensable gases may be injected into the atmosphere. The heavy particles fall out rapidly but the lighter ones injected into the upper atmosphere can linger for years, affecting the absorption of sunlight and the emission of infrared radiation. The consequences for the atmosphere, the ocean and the biosphere can be significant. For example, the dust injected into the atmosphere by the impact of a large meteoritic body may have caused the mass extinctions of organisms at the end of the Cretaceous period 65 million years ago. By the same token, global warming or cooling trends in the more recent past may be correlated with periods of low or high volcanic activity.

One test case for the effect of volcanic aerosols on climate is a global warming of about .4 degree C. that occurred between 1900 and 1940. It should be noted, however, that this global warming is barely significant compared with statistical fluctuations in global temperatures and the bias introduced into the measurements by urban development around many weather-observing stations. Owen B. Toon and James B. Pollack of NASA's Ames Research Center have pointed out that the warming might have been due in part to the clearing of the atmosphere after a period of high volcanic activity from 1880 to 1910. Variations in the global load of atmospheric dust with time can be estimated from a catalogue of eruptions and estimates of the size of each eruption, assuming a stratospheric residence time for particulates of about two years (derived from the dates of reports of intensely red sunsets following the explosion of Krakatau in 1883 and from the fallout rates of radioactive debris from aboveground nuclear tests).

Given an estimate of the total volume of material in the stratosphere and assumptions about average particulate size and composition based on measurements of aerosols currently in the stratosphere, the increase in the amount of sunlight reflected into space and the decrease in the amount of infrared radiation emitted into space can be estimated. The measurements indicate that the effect of the particles on incoming sunlight is greater than their effect on outgoing infrared radiation. The stratospheric aerosols therefore cool the earth. James E. Hansen and his colleagues at NASA's Goddard Institute for Space Studies have used these estimates and those of the contemporaneous increase in atmospheric carbon dioxide to test their climate model. They found that these two factors could indeed have given rise to the observed global warming between 1900 and 1940.

More recent volcanic eruptions have been studied in greater detail, so that more precise estimates of their effect on the climate can be made. It has become clear that the total volume of material ejected by a volcano is not entirely reliable as a guide to its effect on climate; the force of the eruption and the size and composition of the particulates also need to be considered. For example, the eruption of Mount St. Helens in Washington in May, 1980, was larger than that of El Chichón in Mexico in April, 1982, but lofted less stratospheric aerosol. Most of the particles ejected by Mount St. Helens were large and fell out of the atmosphere in a matter of weeks. Although less material was ejected from El Chichón, more of it reached the stratosphere and remained there.

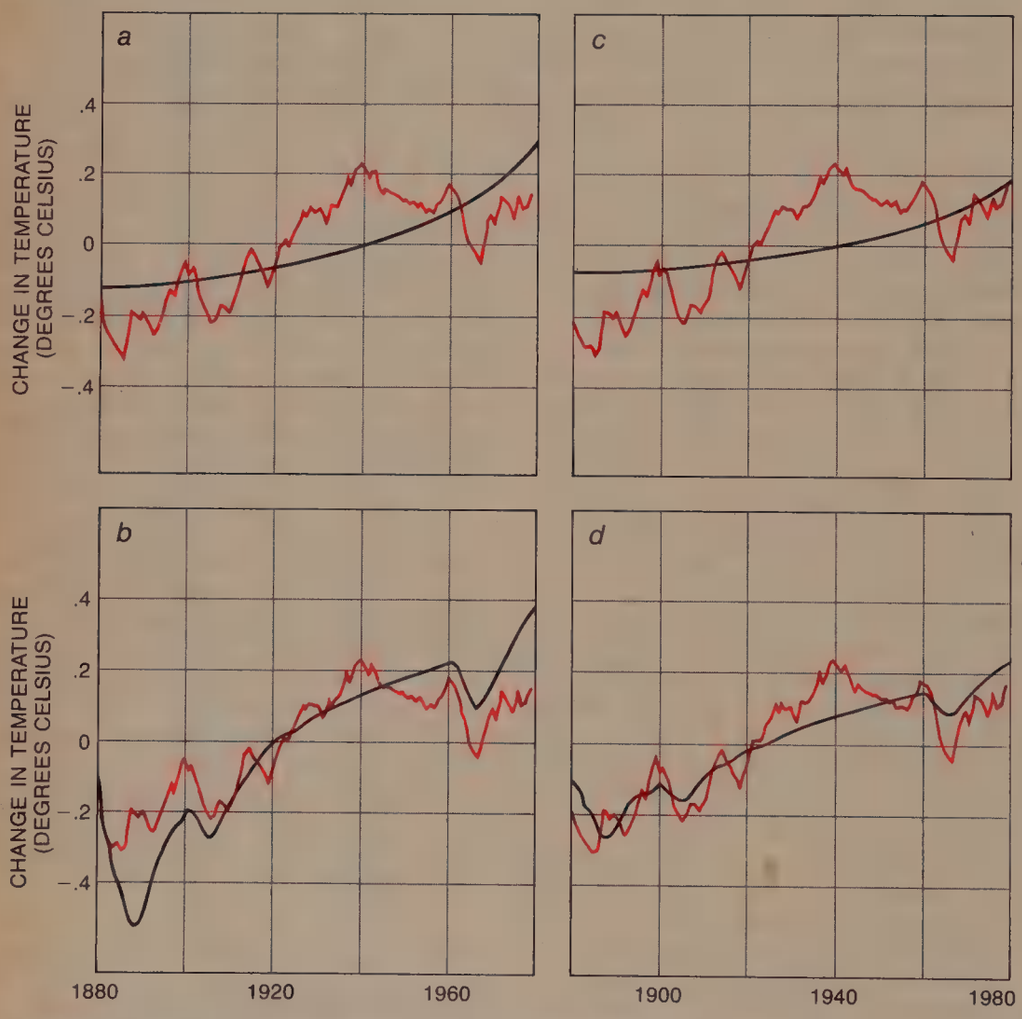
El Chichón apparently emitted a larger quantity of sulfur than Mount St. Helens. The sulfur forms sulfur dioxide, which reacts with water vapor in the stratosphere to produce a smog of sulfuric acid droplets. Such droplets are chemically stable and settle very slowly out of the stratosphere. Predictions of the ultimate effect of the El Chichón dust cloud on the climate over the next few years are still preliminary, but the cloud is generally expected to cause a mean global cooling at the earth's surface of between .3 and one degree C.

Such atmospheric dramas are ones at which the human species is only a spectator, but it will soon be a participant. The first measurable effect of human activity on the global climate will probably be a global warming due to increases in atmospheric carbon dioxide. The amount of carbon dioxide in the atmosphere and the ocean has been increasing since the Industrial Revolution as a result of the combustion of coal and oil and the clearing of forests (both of which lead to the oxidation of carbon and the release of carbon dioxide). The preindustrial value for the amount is not

known exactly, but it has been estimated at between 250 and 300 parts per million. The amount in the air over Mauna Loa on the island of Hawaii has been measured continuously since 1958 by Charles D. Keeling of the Scripps Institution of Oceanography. In the period between 1958 and 1980 atmospheric carbon dioxide showed a trend of increase from 315 parts per million to 336 parts per million.

The increase in atmospheric carbon dioxide is about half the amount of carbon dioxide estimated to have been released; the ocean acts to absorb much of the surplus. The uptake of carbon dioxide by the ocean, however, is almost impossible to measure. Carbon dioxide is not uniformly distributed in the ocean as it is in the atmosphere, so that accurate measurements at one point, such as Keeling's, would not have much meaning. Moreover, the amount of carbon dioxide in all its forms (dissolved gas, bicarbonate ion, carbonate ion and organic carbon) in the oceanic reservoir is much larger than the amount in the atmospheric reservoir, so that the changes are likely to be smaller and harder to detect. Finally, the ocean reacts with carbonate sediments such as limestone, which, at least on a long time scale, may act as an even larger sink of carbon dioxide.

Wallace S. Broecker of the Lamont-Doherty Geological Observatory of Columbia University and others have worked with models of ocean chemistry and circulation to estimate the oceanic carbon dioxide increase from the known atmospheric increase [see "The Ocean," by Wallace S. Broecker, page 146]. The rate of ocean uptake is strongly influenced by the rate of ocean mixing, both the rate of mixing in the surface layer and that of mixing between the surface layer and the deeper ocean. Broecker's model incorporated estimates of the rate of ocean mixing inferred from the dispersal of the radioactive isotopes tritium (hydrogen 3) and carbon 14 manufactured by nuclear-weapons tests. These calculations show that since the Industrial Revolution an amount of carbon dioxide approximately equal to the amount that has remained in the atmosphere has gone into the ocean. The computed total carbon dioxide increase is slightly less than, but nonetheless within the uncertainties of, the estimates of the total amount of carbon dioxide that can be accounted for by the burning of fossil fuels and the clearing of forests. Hence the estimates are at least internally consistent.



CARBON DIOXIDE AND VOLCANIC-AEROSOL LEVELS may be largely responsible for the slight warming trend in global temperature since 1880 (*color*). This period was used as a test case for a climate model (*black*) developed by James E. Hansen and his colleagues at NASA's Goddard Institute for Space Studies. Some versions of the model (*a, c*) considered only the warming due to increasing atmospheric carbon dioxide (assuming that a doubling of carbon dioxide results in a rise in temperature of 2.8 degrees C.). Other versions of the model (*b, d*) also took account of the warming due to the clearing of the stratosphere. The rise in temperature from 1910 to 1940 appears to be due largely to the clearing of the stratosphere during a period of low volcanic activity. The oceans moderate the temperature fluctuations because of their capacity to store large amounts of heat. In some versions of the model (*a, b*) only the heat capacity of the well-mixed surface layer of the oceans was considered. In others (*c, d*) the upper ocean was assumed to mix across the thermocline between it and the deeper ocean.

Armed with these estimates of the partitioning of carbon dioxide between the oceans and the atmosphere and estimates of future fuel combustion, one can project the increase in the atmospheric carbon dioxide into the next century and beyond. According to most projections of future energy consumption, the amount of atmospheric carbon dioxide will double by the middle of the next century and will probably double again before the peak of fossil-fuel consumption is reached.

Mathematical models can then be utilized to predict the consequences of the projected increase for the earth's climate. For example, Syukuro Manabe and Richard T. Wetherald of the Geophysical Fluid Dynamics Laboratory estimate that a doubling of atmospheric carbon dioxide (from 300 parts per million to 600 parts) would increase the mean global temperature by 2.5 degrees C. At high latitudes the temperature increase is likely to be closer to five degrees C.

Fluctuations in local temperature larger than 2.5 degrees occur daily, but such a change in the mean global temperature is significant: the change would be as great as that between the height of the last glaciation 18,000 years ago and today. The secondary consequences of the warming of the earth due to the increase in the carbon dioxide level might include a further melting of the earth's polar ice, a rise in sea level by tens of meters and more precipitation at high latitudes (perhaps offset by a widening of the subtropical arid zones, which would greatly alter the distribution of the world's water resources).

As is often the case, the greatest uncertainties in these estimates are associated with the projections of what the human species will do. This does not mean, however, that environmental scientists will sit back and await the unknown. Rather, as the environmental consequences of the human drama unfold they will keep the world informed and if necessary sound the alarm. Perhaps, on the other hand, by the next century the human species will have learned to control or influence the climate for its own benefit.

We've got a hit on the charts.

One picture is worth a thousand words. Or numbers.

Now there's a simple way to transform pages and pages of business data into a single, dramatic format.

Introducing DR Graph™—high quality business graphics software from Digital Research.™ It lets you create impressive business charts, graphs, or any combination of both—quickly and easily. You can even build graphs from electronic spreadsheets such as VisiCalc® and SuperCalc.™

See what you're creating, instantly.

Think of it as painting by numbers. Because DR Graph actually lets you create your graph step-by-step.

First, enter the data you wish illustrated. Then just choose from the menu, hit a key, and take a quick look at what you've created—instantly.

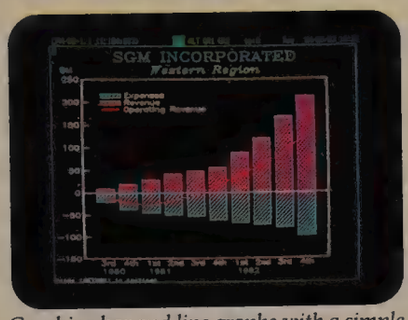
Make your axes thick, your border lines thin. Go with solid, or broken grid lines. Add color when and wherever you please for all your special reports. It's easy to experiment until the graph format is perfect. Then DR Graph can print it, plot it, or store it for future use.

You get the versatility you need.

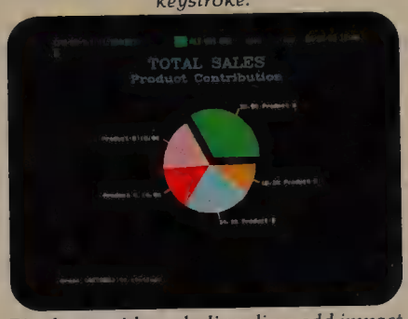
DR Graph lets you convert business data into dozens of personalized choices. Do you like your bars vertical or horizontal, stacked or clustered? How about adding line curves to your bar graph? Or special labels on the axis? With DR Graph, you've got it. All with the simple stroke of a key. DR Graph even displays up to four different graphs on a single page.



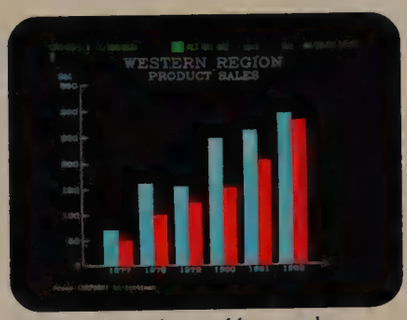
Multiple graphs on a single page



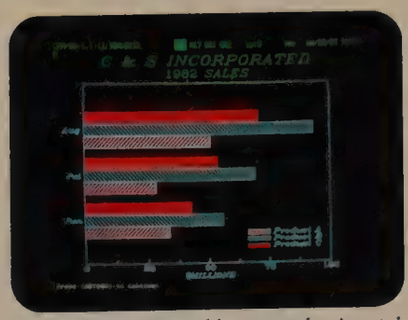
Combine bar and line graphs with a simple keystroke.



Pie charts with exploding slices add impact.



Create clustered bar graphs.



Tap a key and vertical bars turn horizontal.

Mix words with your pictures.

DR Graph lets you write comments on your graphs, too. You not only control the size and color of the type, but also have four different typestyles to choose from. Write your own legends and titles. Or add additional text right onto the graph—exactly where you want it.

Plus, its 8 color and 6 pattern fills provide customized highlighting. So, your presentation graphs will be as professional as you are.

All the business graphics you'll ever need, ready to roll.

DR Graph works with today's leading microcomputers, as well as with a wide range of printers and plotters. All you need is Digital Research's GSX™ graphics enhancement for your computer's operating system. Contact your computer manufacturer, or stop by the CP/M Library™ at your computer shop for an eye-opening demonstration. Call 800-227-1617, ext. 400 (in California 800-772-3545, ext. 400) for a free, full-color brochure.

DIGITAL RESEARCH™

The best of everything in business graphics.

VisiCalc is a registered trademark of VisiCorp. SuperCalc is a trademark of SORCIM Corporation.
The logo, tagline, DR Graph, GSX and CP/M Library are either trademarks or registered trademarks of Digital Research Inc. ©1983 Digital Research Inc.

The Biosphere

The totality of microbial, animal and plant life on the earth not only is sustained by the lithosphere, the hydrosphere and the atmosphere but also has powerfully shaped their evolution

by Preston Cloud

Of all the dynamic systems that make up our planet the biosphere is the last to have taken form, the most distinctive and the most interactive. Other planets and satellites in the solar system have a core, a mantle and even a crust and an atmosphere. Titan, the largest satellite of Saturn, may have seas of liquid methane, and from Mars outward through the solar system there are many icy hydrospheres. Only on the earth, however, are there structures that can replicate themselves, change into different forms by mutation and genetic recombination and transmit such changes to their descendants. Structures that can do these things are said to be living, and it is their integration everywhere on and near the earth's surface with the lithosphere, the hydrosphere and the atmosphere that constitutes the biosphere.

The diversity and interactiveness of the biosphere are rich beyond current understanding, and the potential variety within its collective gene pool is staggering. Biologists have so far catalogued some 1.5 million species of animals and .5 million species of what are generally called plants (algae, fungi and bacteria included). New species are being found and named at the rate of some 10,000 yearly. Most of the new species are insects, but the number includes such other living forms as giant deep-sea clams and numerous fossils from the still undescribed legions of plants and animals that flourished in earlier epochs. Indeed, presented with the least excuse, life teems. Responding to the variety of the earth's potential ecological niches, organisms have in geologic terms often diversified rapidly. Here we see evolution at work. Climatic and geographic isolation and such grand-scale variables as motions of the earth's lithospheric plates and the intensity of the sun's radiation provide the selective pressures that give direction to a process that otherwise seems random.

Consider the Hawaiian Islands, generated by the drifting of a lithospheric plate over a "hot spot," a rising convection current in the mantle. Kauai to the northwest, formed as the drift began, is 5.6 million years old, Oahu is 3.3 million and Maui is 1.8 million. At the southeast end of the chain the big island, Hawaii, which is still growing, emerged only 700,000 years ago. In that brief 5.6 million years of earth history the descendants of some random fruit-fly migrants to the new islands have so diversified that 25 percent of the world's total of some 2,000 species of *Drosophila* are found there and there alone. Over the same span in the island chain more than 1,000 separate species of land snails evolved in the isolated valleys that descend the volcanic slopes. And an entire new family of birds—the honeycreepers, numbering some two dozen species—has arisen from unknown finchlike ancestors, borne to the islands on some chance wind.

In the long history of the biosphere, however, the most important event by far was the origin and early diversification of life: the beginnings of biospheric interaction with the earth's surface. Indeed, I shall not be much concerned here with the fascinating diversity and complexity of the planet's living structures or with their detailed evolution. The focus instead will be on the far-reaching ways in which life has modified the history and the surface features of our world and has in turn been modified over the four billion years that have elapsed since the earth's surface first became habitable.

Today few places on the earth's surface, or below it for some tens of meters, or above it to a height of nearly 10 kilometers, are now so hot or so cold or so dry or so polluted or so exposed to ionizing radiation that they are truly devoid of life. This was not always the case. Nor was the major work of the biosphere done by its currently most conspicuous constituents: the familiar multicellular plants and animals of field, stream and sea. In the course of the biosphere's history its geochemically most influential members have always been those of the microbial realm, the humble and morphologically simple but biochemically diverse and adaptable masses of the minute.

The interaction of the biosphere, the hydrosphere and the atmosphere with one another and with the earth's outer crust is not only universal but also continuous, a matter of cycles within cycles within cycles. Animals such as worms, ants, moles and gophers till the soil until their own remains are recycled by decay bacteria and other scavengers. Meanwhile other biospheric agents, including such odd animal-plant symbionts as termites and their hind-gut bacterial helpers, recycle much of the plant kingdom, simultaneously adding large inputs to the methane and carbon dioxide components of the atmosphere. Every year living plants transfer to the atmosphere some 400,000 tons of organic volatiles, some of them incorporating metals.

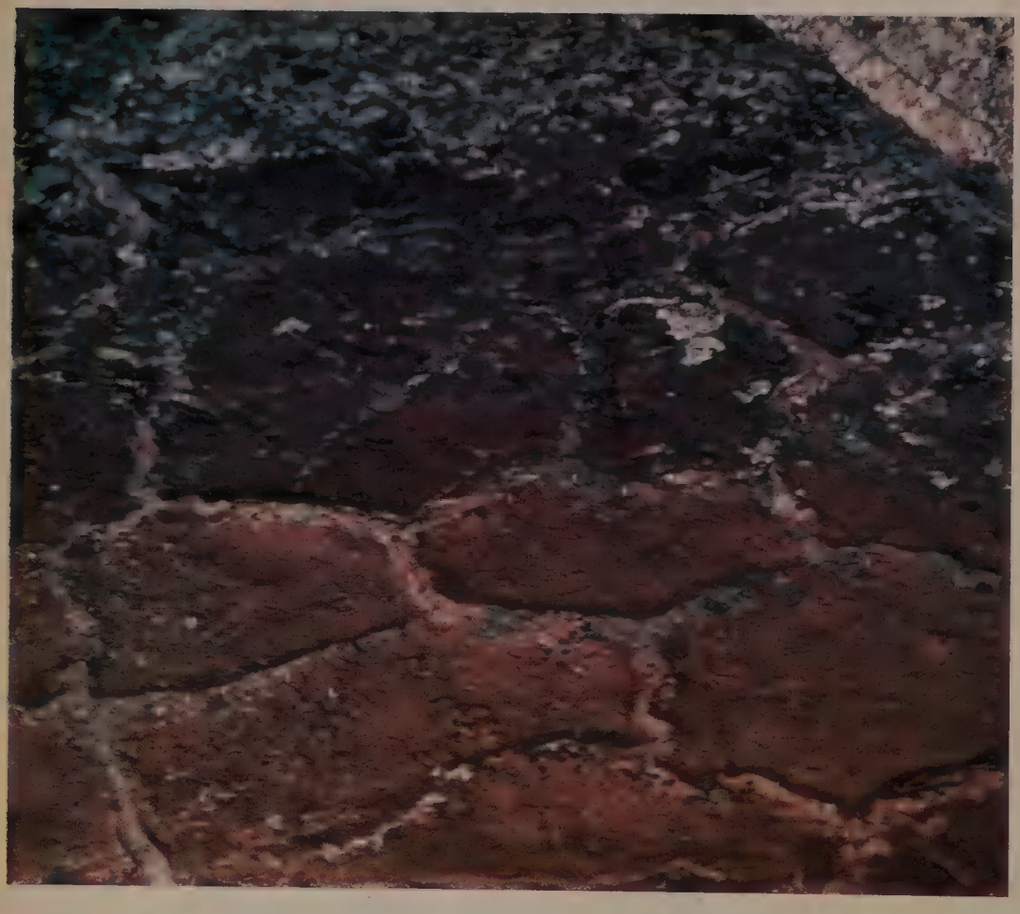
A good third of all chemical elements are recycled biologically, and some living organisms are capable of concentrating comparatively rare elements far above their background abundance. All organisms utilize and recycle not only hydrogen, oxygen, carbon and nitrogen but also phosphorus, sulfur, calcium, potassium, magnesium, sodium, iron, manganese, cobalt, copper and zinc.

GREAT BARRIER REEF, in the vicinity of Cape Melville (*upper left*) in northeastern Australia, is seen in the Landsat image on the opposite page. The reef is a vivid example of how processes of the biosphere interact on a large scale with those of the lithosphere, the hydrosphere and the atmosphere. The activities of diverse life forms, notably corals and algae, have built this 2,000-kilometer natural breakwater extending from the Tropic of Capricorn northward to Torres Strait off New Guinea. The reef parallels the Queensland coast at distances ranging from 15 to 150 kilometers, and the waters between the outer reef and the shore hold many other reef structures. North is toward the top. The area shown is 60 kilometers wide.





OLDEST SEDIMENTARY ROCK, from the Isua area in southwestern Greenland, was deposited some 3.8 billion years ago. The specimen is from what is called a banded iron formation (although it lacks the reddish color found in the alternating dark stripes of other such rocks). The stripes are thinly stratified layers, now folded and stretched. Formations of this kind were among the oxygen "sinks" that kept free oxygen nonexistent or scarce in the early atmosphere of the earth. The rhythmic banding of the specimen, together with a ratio of carbon isotopes reminiscent of life processes, suggests that microbial organisms may already have been at work.



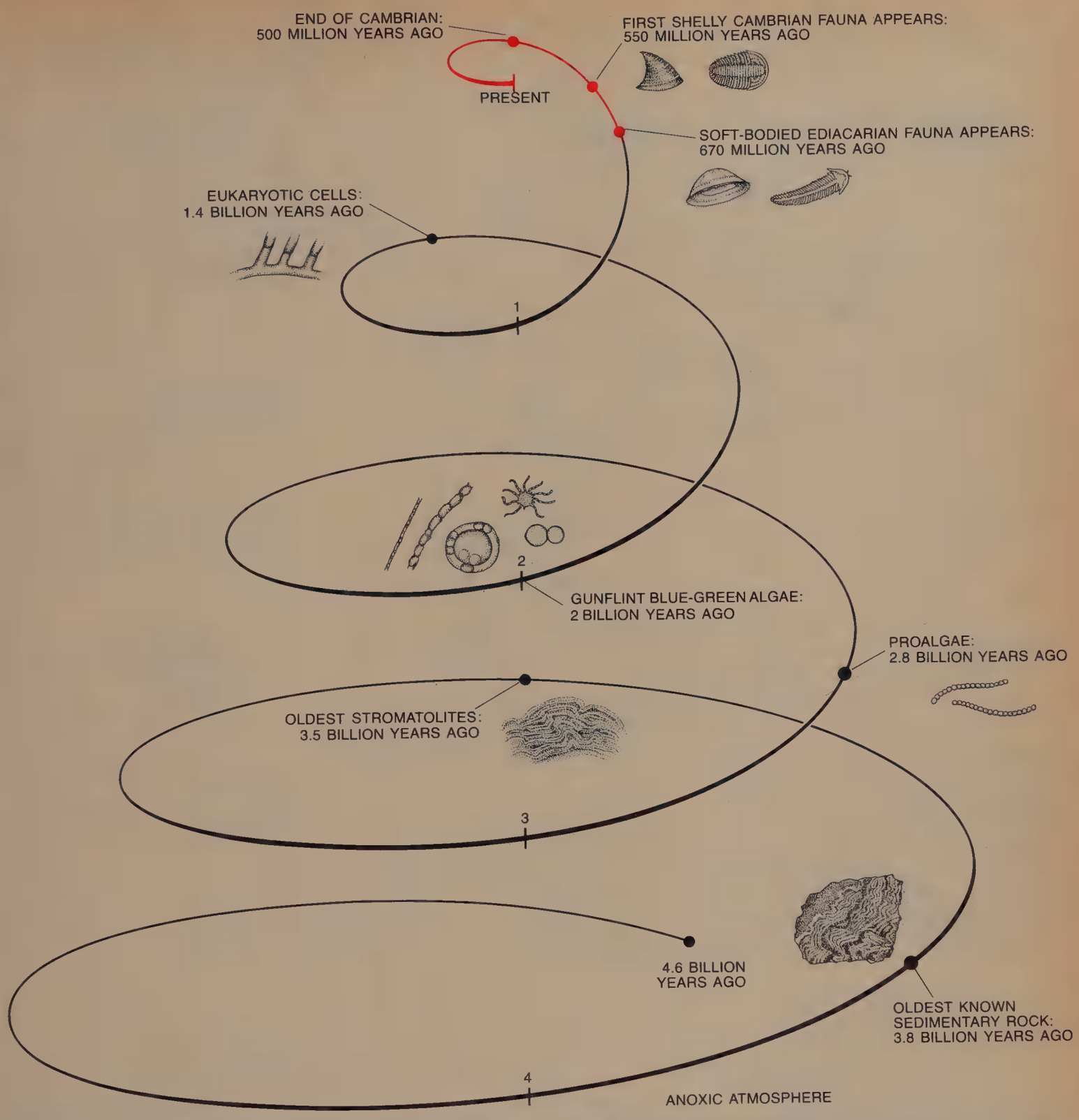
RED SHALE, estimated to be about two billion years old, overlies Cambrian granites near Lac Cambrien in Quebec that are roughly 2.5 billion years old. Among the oldest "red beds" known, it underlies the youngest banded iron formations. Red beds mark a transition from essentially anoxic conditions to the inauguration of an atmosphere where some oxygen was always present.

Most organisms utilize chlorine in some way, and about a dozen other elements are exploited for particular biological functions. Biological processes are responsible for massive concentrations in the earth's crust of silicon, iron, manganese, sulfur and carbon. Microbes can thrive on or in the presence of such corrosive substances as sulfuric acid, carbolic acid and hydrogen sulfide.

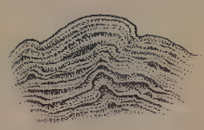
Rock weathering, soil formation and many sedimentary rocks are important partial products of microbial and other biological processes. The oxygen in the atmosphere is overwhelmingly from photosynthesis by green plants, which also segregate carbon in soils and sediments. The domestic animals that provide amino acids essential to human nutrition would not be able to function without their digestive bacteria, nor would we. And many of the plants on which animals feed depend on the nitrogen-fixing activities of bacteria and their first cousins the blue-green algae or proalgae—carryovers from the earth's earliest history. The biosphere profoundly and pervasively affects the rest of the earth's surface and is in turn acutely responsive to feedbacks from other spheres of activity.

The appearance of life was thus a dramatic innovation on the barren prebiotic planet, introducing an unprecedented and far-reaching geologic process. Although one cannot say with confidence just how and when life first arose on the earth, conjecture is now limited by evidence on likely abiotic sources for the initial organic but nonliving macromolecules and the autocatalytic processes that might have organized them into components of the earliest cells. What can be said with some confidence is that the planet's initial life forms were biochemically simple, unicellular (or noncellular), probably spheroidal in shape and dependent on extracellular sources of nutrients. If such objects were eventually to be found as fossils, they probably could not be distinguished on morphological grounds from similar objects of nonbiological origin.

It is also possible to estimate within limits when and under what conditions the protobiosphere first began and even to reconstruct some elements of its subsequent history and interactions with the nonbiological world. The evidence on biospheric processes in earth history belongs to the broad field of biogeology. Its data include the remains of the microorganisms found, sedimentary structures produced by their activities, the geochemistry and isotopic composition of biologically significant elements and the biogeochemistry of organic substances and products. The oldest two major signposts are both in Western Australia. One of them, near a place so isolated that it is called North Pole, is



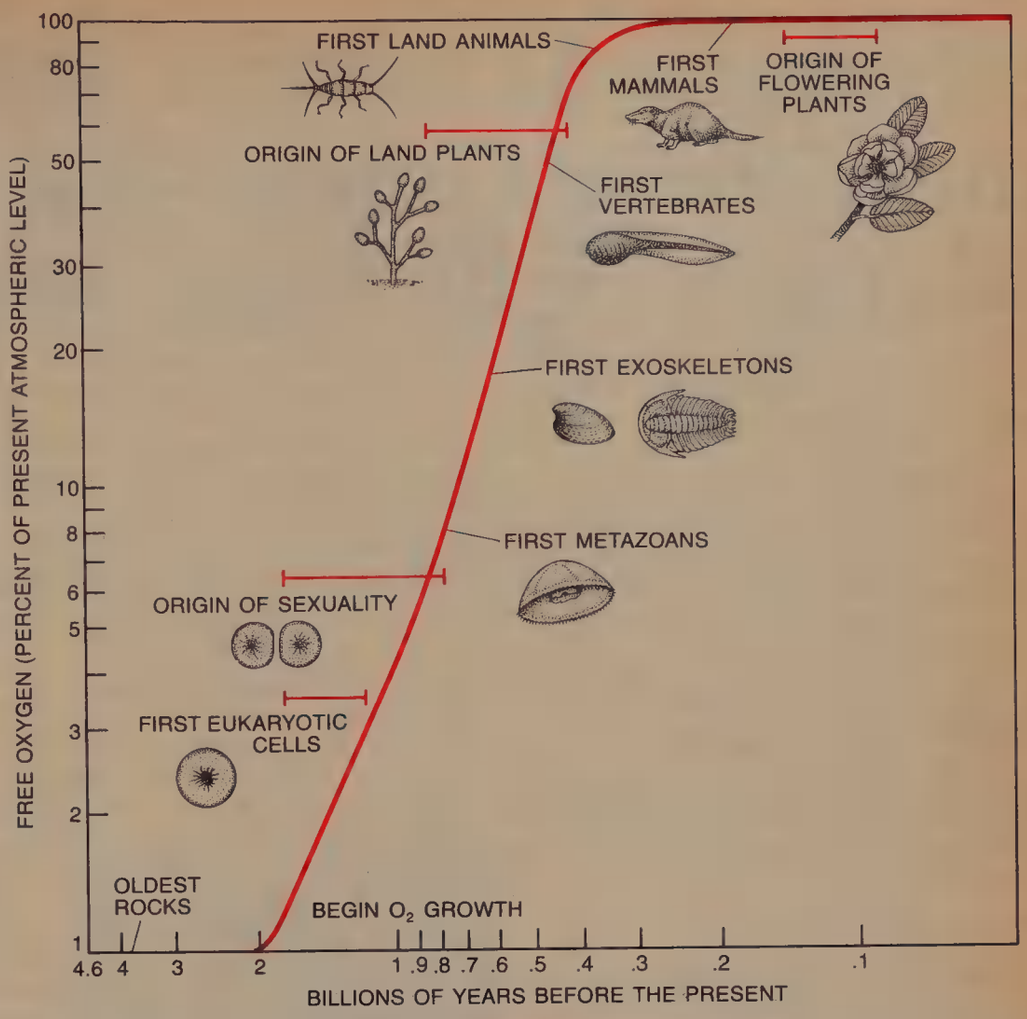
ASCENDING SPIRAL schematically represents the earliest phases of the earth's biogeologic history, beginning soon after the solar system formed some 4.6 billion years ago, when the planet was inhospitable to life and a direct geologic record is nonexistent. By 3.8 billion years ago the developing atmosphere probably consisted mainly of carbon dioxide, water vapor, nitrogen, carbon monoxide, hydrogen sulfide and hydrogen. Free oxygen was absent or trivial and transient. The temperature of the earth's solid surface had by then fallen from about the melting point of iron to a mean temperature between the boiling and freezing points of water. Even though the young sun shone with only 60 to 70 percent of its present intensity, a life-supporting range of temperatures was made possible by the "greenhouse effect" of an atmosphere rich in carbon dioxide. Some 300 million years later sedimentary structures called stromatolites appear; they resemble deposits formed today by the photosynthetic and sediment-binding activities of blue-green algae. After about another 700 million years sedimentary rocks being deposited in what is now Western Australia included tiny filamentous structures that are convincingly microbial and may be proalgae. By two billion years ago, a date on the third loop of the spiral, accumulating silica gels of a Lake Superior iron formation were trapping filaments, cells and more complex structures that are unquestionably microbial. Some show two or three cell types similar to the living blue-green alga *Nostoc*, a typical prokaryote. (Prokaryotic cells lack the nucleus characteristic of eukaryotic cells, the building blocks of all higher forms of life.) By then the oxygen content of the hydrosphere and the atmosphere is estimated to have reached about 1 percent of its present level, allowing a tenuous ozone screen to develop and shield the earth's surface from ultraviolet wavelengths inimical to living organisms. The oldest persuasively eukaryotic cells entered the fossil record about 1.4 billion years ago. Their appearance was announced by a widespread increase in the average diameter of the cells. By about 670 million years ago a remarkable and diverse group of soft-bodied aquatic animals, the first known metazoans, entered the fossil record in the sedimentary deposits of the Ediacara Hills in South Australia. Their lack of a shelly covering suggests that the oxygen level was by then near 7 percent of the present level. Their appearance marks the end of earlier divisions of geologic time and the start of the present eon, the Phanerozoic. This is the period of manifest rather than cryptic animal life that appears in color on the fifth and final loop of the spiral. Soft-bodied and shelly metazoans are preserved as fossils in Cambrian sediments, deposited between 550 and 500 million years ago.

EVENT	BILLIONS OF YEARS AGO	MANIFESTATION	OXYGEN (PERCENT)	COLLATERAL EVENTS AND CONSEQUENCES
8. FULLY OXIC CONDITIONS	.4	 LARGE FISHES, FIRST LAND PLANTS	100	BIOSPHERIC EVOLUTION TOWARD PRESENT WORLD STATE
7. SHELLY METAZOANS APPEAR	.55	 CAMBRIAN FAUNA	~10	BURROWING HABIT BEGINS; SUBSEQUENT EVOLUTION
6. METAZOANS APPEAR	.67	 EDIACARIAN FAUNA	~7	PHANEROZOIC EON BEGINS; METAZOAN FOSSILS AND TRACKS
5. FIRST EUKARYOTIC CELLS FOUND	1.4	 CELLS LARGER IN DIAMETER	>1	RED BEDS PROLIFERATE, MUTICELLULAR ORGANISMS; MITOSIS, MEIOSIS, GENETIC RECOMBINATION
4. OXYGEN-TOLERATING BLUE-GREEN ALGAE	~2.0	 ENLARGED, THICK-WALLED CELLS AT INTERVALS ON ALGAL FILAMENTS	~1	OXIDATIVE METABOLISM, OZONE SCREEN; OLDEST RED BEDS OVERLAP WITH YOUNGEST BANDED IRON FORMATIONS
3. PHOTOAUTOTROPHY, PROBABLY RELEASING MOLECULAR OXYGEN	>2.8	 STROMATOLITES, PRECURSORS OF BLUE-GREEN ALGAE	<1	CHLOROPHYLL *a* AND CYTOCHROME *b* SYNTHESIZED; BANDED IRON FORMATIONS AND OTHER OXYGEN SINKS; GREENHOUSE EFFECT DIMINISHED
2. AUTOTROPHY (METHANOGENESIS?) (SULFUR OXIDATION?)	>3.5	 STROMATOLITES, SULFATE, LIGHT CARBON	(ANOXIC)	BIOSPHERIC CONTINUITY
1. ORIGIN OF LIFE	(~3.8?)	 LIGHT CARBON	(ANOXIC)	EVOLUTION OF BIOSPHERE BEGINS

EIGHT MAJOR STEPS in the early evolution of the biosphere are listed in this table, beginning at the bottom with the origin of life on the earth about 3.8 billion years ago (*1*). This date is suggested by the diminution of the isotope carbon 13 with respect to the lighter isotope carbon 12 ("light carbon" in the table), a ratio characteristic of life processes. A few hundred million years later (*2*), although oxygen was not yet permanently present in either the atmosphere or the hydrosphere, the appearance of stromatolites implies that some microbial form of self-feeding organism had evolved. Perhaps 800 million years after that (*3*) additional stromatolites and the fossilized remains of what appear to be blue-green algae or their precursors lead to the conclusion that oxygen-releasing photosynthesis might have begun, presenting the problem of how life was to be shielded from this corrosive element. The appearance of banded iron formations, whose sediments served as large oxygen sinks, may have postponed the problem. About two billion years ago (*4*) blue-green algae whose chains of tiny cells included thick-walled ones such as those shielding nitrogenase enzymes from free oxygen today indicate that oxygen was accumulating in the hydrosphere and suggest the appearance of oxidative metabolism as a superior energetic pathway. The first continental red beds appear at about this time, implying development of an atmosphere that contained oxygen permanently. Between then and 1.4 billion years ago (*5*) eukaryotes, generally larger cells whose DNA is shielded within a membrane-bounded nucleus, signal the evolution of three characteristic traits: mitosis (in which rodlike chromosomes line up and divide in a spindle), meiosis (in which the chromosomes are divided in half for reproductive purposes) and the advanced evolutionary process of genetic recombination. Between 670 and 550 million years ago fossils found at sites on five continents (*6*) present the outcome of the continuing proliferation of eukaryotes: the Ediacarian fauna, multicellular aquatic animals that mark the start of the Phanerozoic eon, in which all life on the earth still lives. By 550 million years ago (*7*) the Cambrian fauna, a worldwide succession of ancient organisms, including the first (or almost the first) shelled invertebrates, enters the fossil record. It is estimated that by then the supply of oxygen had reached about 10 percent of the current level. Some 150 million years later (*8*) the probability that the present level of oxygen was being approached is implied by the presence of large and active fishes in the sea and plants and invertebrate animals on the land. Their descendants went on to fill all the planet's ecological niches.

found in rocks about 3.5 billion years old (the Warrawoona Group) in the form of inferentially biogenic sedimentary structures called stromatolites (and microfilaments that may not, however, be contemporaneous with the enclosing sediments). The second signpost, from 2.8-billion-year-old strata of the Fortescue Group, also in Western Australia, contains the oldest fossil paleomicrobiota that I find to be convincing both as to their biogenic nature and as to their contemporaneity with the rocks in which they are found.

A third signpost is firmly planted in two-billion-year-old siliceous rocks called chert on the north shore of Lake Superior. They are a part of the Gunflint Iron Formation, which contains a variety and abundance of demonstrably biogenic and unequivocally contemporaneous microbial fossils that compare favorably with the richest modern microfloras. Indeed, the Gunflint microbiota contains the oldest fossils known to display a clear differentiation into two or more types of cell. A common filamentous microfossil closely resembles living freshwater blue-green algae of the genus *Nostoc*. Others resemble budding bacteria. Their presence in this formation, together with their resemblance to living organisms, implies a continuity of biological functions from two billion years ago to today.



ENRICHMENT IN OXYGEN of the earth's originally anoxic atmosphere is traced on this double logarithmic graph. The abscissa measures billions of years before the present; the ordinate shows the estimated increase of molecular oxygen from about 1 percent of its present level to what it is today. Leaders indicate the timing of key events in the evolution of the biosphere as they relate to the evolution of the atmosphere. In the case of events where there is uncertainty about the time of origin a range bar (*color*) indicates the possible limits. For example, microbial land plants of the simplest kind might have appeared at or before the beginning of the Phanerozoic eon, but spores that have been interpreted as being those of true vascular plants did not appear until the late Ordovician, some 440 to 430 million years ago. In turn, the oldest organisms generally accepted as being land plants in the modern sense are of late Silurian age, perhaps 420 to 415 million years old. Scorpionlike arachnids of about the same age are known. Primitive insects do not appear until the Middle Devonian, some 380 million years ago.

Earlier records of microbial life are less conclusive. The record of the 2.8-billion-year-old Fortescue Group consists of filamentous chains of cell-like entities that show no cellular differentiation but otherwise resemble certain living blue-green algae. The 3.4-to-3.5-billion-year-old Warrawoona rocks at the locality from which the best material was collected show three separate sets of postdepositional fracturing, have undergone changes in their initial chemical state and are heavily infiltrated by secondary iron. The best-preserved and most abundant microbial forms in these rocks resemble the flat, twisted stalks of the modern iron bacterium *Gallionella ferruginea*. Evidence that they are the same age as the sediments in which they are found is not convincing. In rocks that are considered to be the same age as those containing this microflora, however, there are laminated and undulating pseudocolumnar and domal stromatolite structures like those being built in shallow waters today by colonies of blue-green algae. These very old stromatolites are presumptive evidence (not proof) of a microbial presence in Warrawoona time, perhaps the progenitors of later blue-green algae or their bacterial cousins.

Even more indirect evidence for a still earlier signpost is found in carbonaceous sediments of the Isua area in southwestern Greenland. The carbon-isotope ratios of these 3.8-billion-year-old rocks show a depletion of the isotope carbon 13 with respect to the lighter isotope carbon 12, a common manifestation of biospheric activity. It would be consistent with such evidence to postulate the presence of life even that early. The fractionation observed, however, is not conclusive as to its source. It could be the work of aerobic or anaerobic photosynthesizing organisms, of methane-assimilating bacteria (methanogens) or conceivably of some nonbiological process.

What was the terrestrial environment like in those early times? So far no rocks have been found on the earth older than those at Isua, although much older rocks are known on the moon and in the form of stony meteorites. Considering the evidence for a high rate of meteoritic and cometary bombardment throughout the early history of the solar system, however, it can safely be inferred that up to about four billion years ago the earth was inhospitable to life as we know it. Initial surface temperatures near or above the melting point of iron, the absence of both an atmosphere and a hydrosphere and intense, unshielded solar radiation would all have contributed to the harshness of the primordial environment. Like life, however, the earth evolves, and these harsh conditions changed.

A free interpretation of the geologic record in the light of basic biological principles suggests the following surface milieu some 3.8 billion years ago. The developing atmosphere at that time was anoxic: it lacked any permanent free oxygen. Its main gases were most likely carbon dioxide, nitrogen, water vapor, carbon monoxide and perhaps hydrogen sulfide. Traces of hydrogen, hydrochloric acid, ammonia and methane would probably also have been present. There were no primary sources of oxygen, neither ordinary molecular

Medicine is always changing, now there's a text that changes right along with it.

SCIENTIFIC AMERICAN *Medicine.*

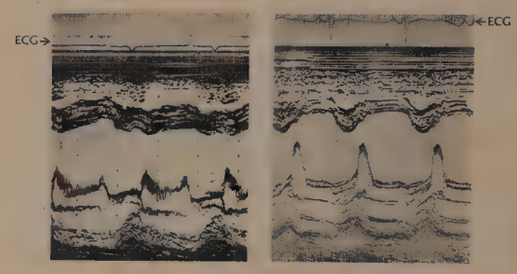
The answers that you get from SCIENTIFIC AMERICAN *Medicine* are current, because each month new chapters replace the old. You will know where to look for the answers, because this advanced text follows the familiar structure of medical knowledge, changing as medicine changes.

Comprehensive. Leading authorities from Harvard and Stanford cover the full range of medicine for you the seasoned clinician.

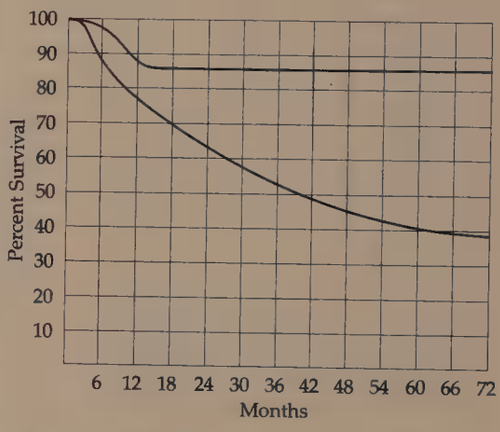
Complimentary CME. At no extra cost, we offer a CME program in which you can earn 32 top credits* per year.

Trial Offer. We invite you to try SCIENTIFIC AMERICAN *Medicine* for two months with our compliments. Send us the coupon and you will receive the two-volume text and two monthly updates. You may also take a CME test for credit. At the end of 60 days, if you decide to continue the subscription, you will owe us $220 for the first 12 months (current renewal is $170); otherwise, return the books; there is no charge.

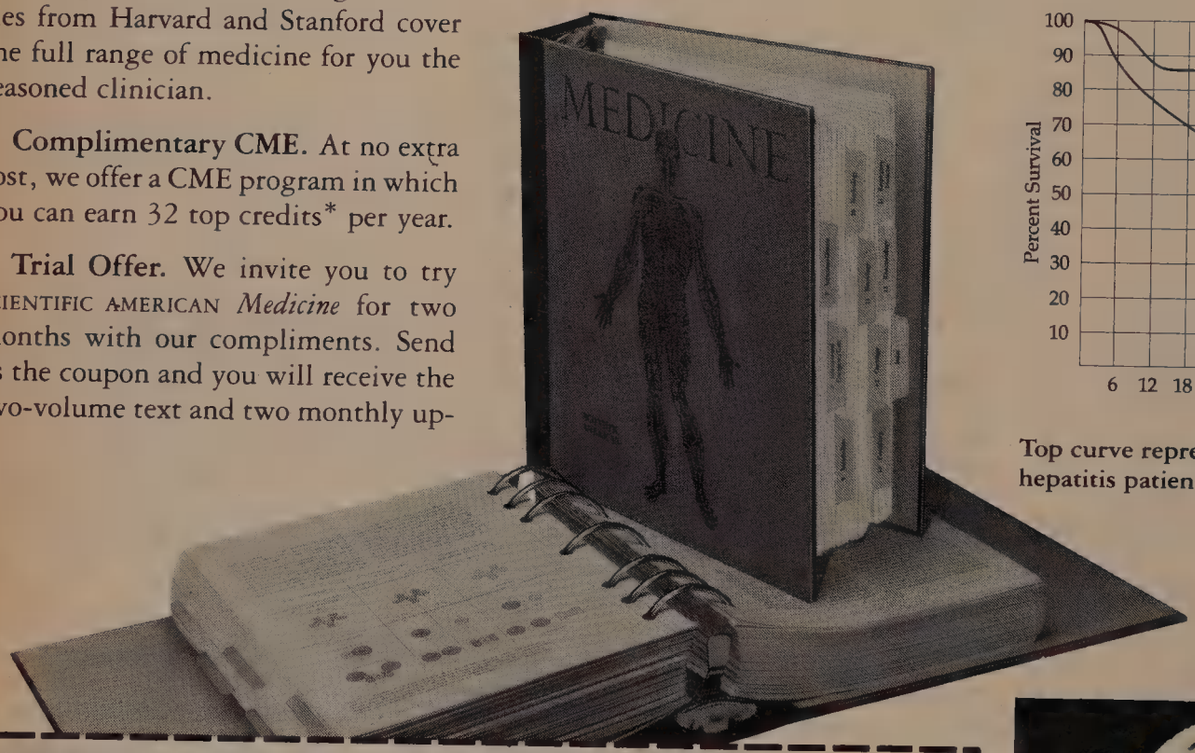
Please mail the coupon today so that we can put medical progress to work for you.



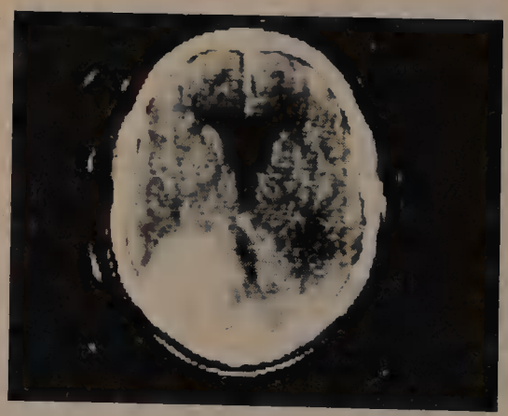
Echocardiograms from patients with aortic regurgitation.



Top curve represents survival in chronic hepatitis patients treated with steroids.



CT scan reveals large meningioma and mild hydrocephalus.



SCIENTIFIC AMERICAN MEDICINE

415 Madison Avenue, New York, New York 10017

Please enroll me as a subscriber to SCIENTIFIC AMERICAN *Medicine*. On receipt of this coupon you will send me the advanced two-volume text described in your announcement and update it regularly by sending me new monthly subsections. I understand that the annual subscription of $220 for SCIENTIFIC AMERICAN *Medicine* is tax deductible. If I am not entirely satisfied, I may cancel at any time during the first 60 days, returning all materials for a *complete refund*.

☐ Please enter my subscription for SCIENTIFIC AMERICAN *Medicine*
☐ I shall also enroll in the CME program
☐ I enclose a check made out to SCIENTIFIC AMERICAN *Medicine* for $220*
☐ Bill me ☐ VISA ☐ MasterCard

Expiration Date_____ Account Number_____

*Please add sales tax for California, Illinois, Massachusetts, Michigan, Ohio and New York

Name_____ MD Specialty_____

Address_____

City_____ State_____ Zip_____

Signature_____

Please allow six to eight weeks for delivery. All payments must be in U.S. dollars. CME is available outside the U.S. and its possessions for an additional $65. Surface routes used unless airmail delivery is prepaid.

7D

You may also order your trial copy by calling toll-free 1-800-345-8112 (in Pennsylvania, 1-800-662-2444).

*As an organization accredited for continuing medical education, the Stanford University School of Medicine designates this continuing medical education activity as meeting the criteria for 32 credit hours in Category 1 for Educational Materials for the Physician's Recognition Award of the American Medical Association, provided it has been completed according to instructions.

This program has been reviewed and is acceptable for 32 prescribed hours by the American Academy of Family Physicians.

This program has been approved by the American College of Emergency Physicians for 32 hours of ACEP Category 1 credit.

oxygen (O_2), singlet or atomic oxygen (O) nor ozone (O_3). Whatever transient free oxygen may have formed as a result of the photolytic breakdown of water vapor and carbon dioxide was rapidly consumed by chemical recombination, by reduced gases expelled from the earth's interior and by a variety of other oxygen sinks. The developing hydrosphere was probably salty, although probably less salty than modern seas as a result of dilution by water of cometary origin. The ancient sedimentary rocks of Isua ensure that the mean surface temperature of the planet as far back as we have any record was above the freezing point of water and below the boiling point. The freezing suggested by the astronomers' "faint early sun" was presumably counteracted by the "greenhouse" effects of atmospheric carbon dioxide, water vapor and ammonia.

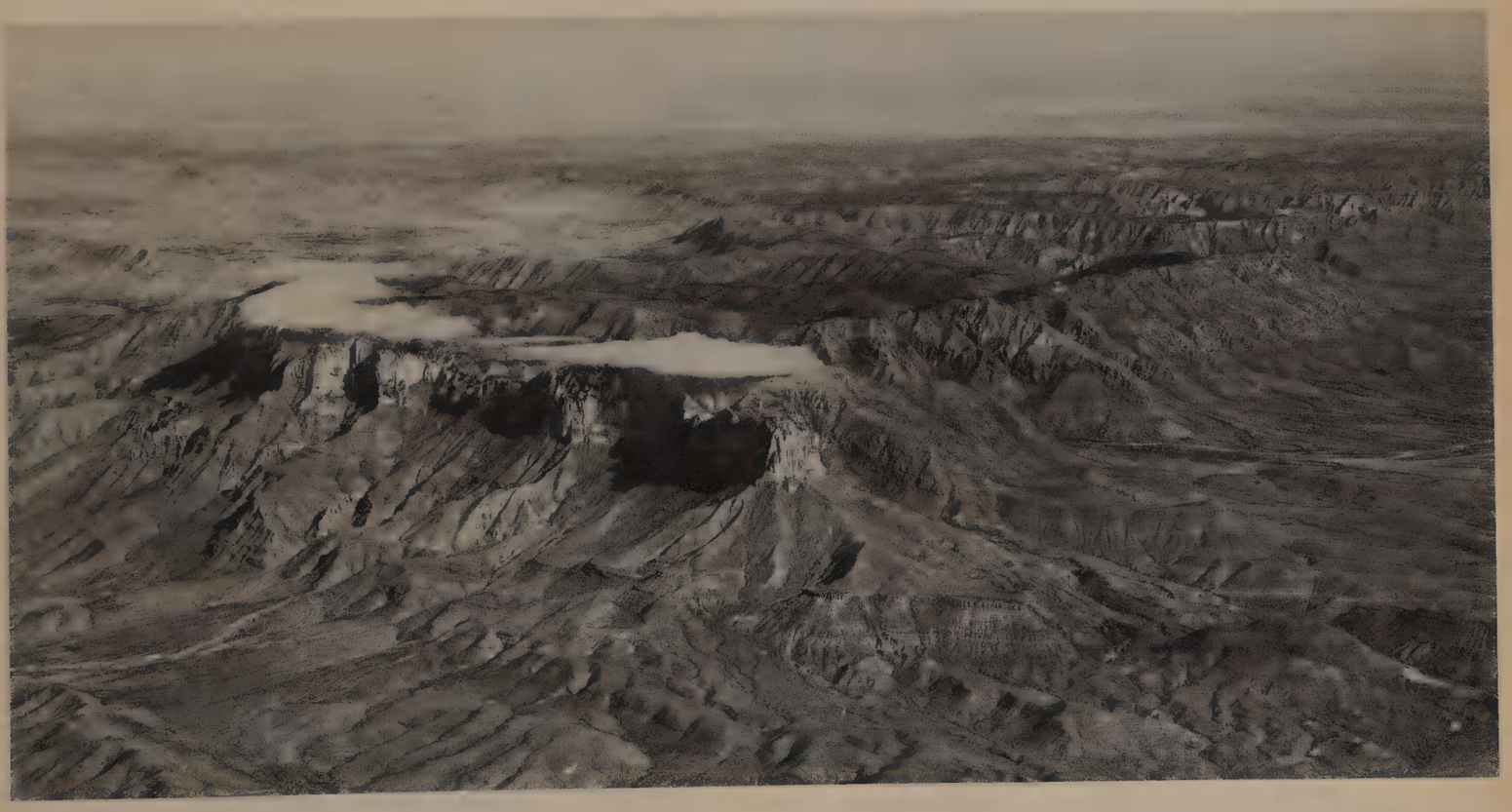
How could life have arisen in such a setting? That calls for a source or sources of energy, a templating structure or mechanism to explain the chirality, or asymmetric "handedness," of amino acids and sugars, a local concentration of organic macromolecules and catalytic effects to hasten and direct the process. Experiment, observation and reflection support the view that energy for prebiotic and early biochemical reactions was available from solar ultraviolet radiation and possibly other sources (for example lightning or chemical reactions). The templating and concentration of abiogenic organic molecules may have been favored by the properties of asymmetric minerals such as clays, or perhaps by polarized light. And autocatalytic effects were surely possible in the form of the frequent day-night freeze-thaw cycles of the rapidly rotating early earth, of iron and magnesium compounds, of dehydration condensations (the linking of molecules by the loss of combined water) and of chemical selection. Whatever the details may have been, the origin of life was an epic event. Given the record of biospheric interactions, nothing at or near the earth's surface could ever have been the same again.

From such an unpromising, oxygen-depleted start how might the initial biosphere have become the biosphere of today? The central and most general driving force was natural selection: a response to changing ecological challenges and opportunities. Among these I shall stress the role of oxygen.

Oxygen, primarily in its molecular form, was to the early biosphere what, by analogy, nuclear power is to the biosphere of today: pregnant with potentialities and cursed with contradictions. The biosphere is a huge metabolic device for the capture, storage and transfer of energy. It carries out these metabolic functions in two modes: fermentation and respiration, both involving the enzymatic conversion of glucose into energy through intermediate steps. One group of organisms, all bacterial, operates by fermentation only. The metabolism of most other organisms, mainly the ones we call higher but also some microbes, is respirational. To the initial fermentative step called glycolysis, respirational metabolism adds the set of reactions known as the citric acid cycle, which through the process of oxidative phosphorylation increases the amount of energy generated 16-fold. In the citric acid cycle glucose is converted into biological energy in the form of adenosine triphosphate (ATP).

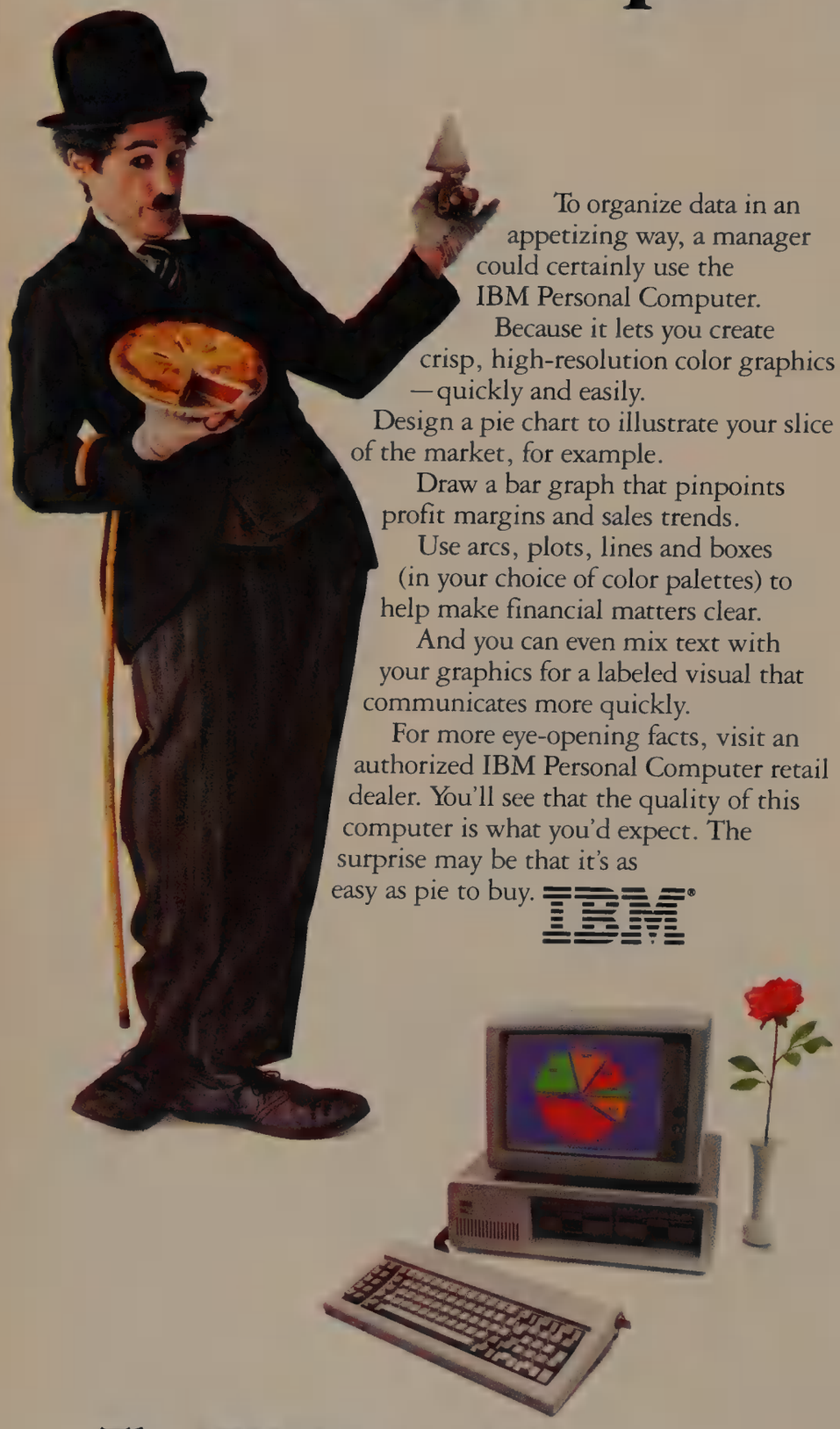
From this it is inferred that fermentation is the primitive form of metabolism and respiration the derived form. Only thus can one account for their near-correspondence, with the separation of life into two primary categories, successional in time and built from dramatically different cell types: (1) the prokaryotes, which lack a cell nucleus and are mainly fermentative, and (2) the eukaryotes, which have a nucleus and are mostly obligatory aerobes. So also can we best account for the superimposing of the citric acid cycle on fermentational beginnings. The blue-green algae are functionally intermediate prokaryotes, some of which can switch off oxidative phosphorylation and function on fermentation alone. The first green-plant photosynthesizers, they have not completely cut their ties with their anaerobic bacterial ancestors.

The principal driving forces behind the interaction of the biosphere with the lithosphere, the hydrosphere and the atmosphere are therefore the production of glucose, utilizing external sources of energy, and the metabolic conversion of glucose into ATP, particularly as enhanced by oxidative phosphorylation in the citric acid cycle. When the early



GUADALUPE MOUNTAINS of Texas, near the New Mexico border, are capped by a great bluff, here covered by clouds, that was built mainly by sponges and algae. It is part of a reef, 640 kilometers long, that gradually arose in late Permian times, 240 to 230 million years ago. Most of this Paleozoic reef is now buried under younger sediments. The cloud-capped peak, El Capitan, reaches 4,000 feet.

How to slice a pie.



To organize data in an appetizing way, a manager could certainly use the IBM Personal Computer.

Because it lets you create crisp, high-resolution color graphics—quickly and easily.

Design a pie chart to illustrate your slice of the market, for example.

Draw a bar graph that pinpoints profit margins and sales trends.

Use arcs, plots, lines and boxes (in your choice of color palettes) to help make financial matters clear.

And you can even mix text with your graphics for a labeled visual that communicates more quickly.

For more eye-opening facts, visit an authorized IBM Personal Computer retail dealer. You'll see that the quality of this computer is what you'd expect. The surprise may be that it's as easy as pie to buy.

IBM®

The IBM Personal Computer
A tool for modern times

For more information on where to buy the IBM Personal Computer, call **800-447-4700.** In Alaska or Hawaii, **800-447-0890.**

blue-green algae or proalgae first left unmistakable traces in the form of the Gunflint microbiota almost two billion years after life may have originated on the earth, the atmosphere probably had only about 1 percent of its present level of molecular oxygen. This meant, however, that the initial oxygen sinks were becoming neutralized and that oxygen was now accumulating in the hydrosphere and beginning to escape into the atmosphere. Oxidative phosphorylation was surely by then a fact of life.

It is now appropriate to consider more fully how the evolution of the biosphere may have been linked with the other great spheres of activity that make up the dynamic earth. Although the evidence is still dauntingly incomplete, it is sufficient to formulate a consistent preliminary account of probable early linkages among these spheres. In addition to the increase of oxygen in the hydrosphere and the atmosphere I shall briefly consider the evolution of the eukaryotic cell, foreshadowing the development some 700 million years later of the Metazoa: multicellular animal life. Thereafter plate tectonics had a profound influence on the evolution of the Metazoa and land plants. Here, however, I shall emphasize microbial processes, later assisted by the higher algae and land plants.

What really needs explaining about atmospheric oxygen is not its scarcity in the beginning but its abundance in the end; from anoxic in the beginning, to perhaps 7 percent of the present atmospheric level when metazoans first appear, to near the present level by mid-Paleozoic times. This much, at least, is clear: omitting minor amounts of photolytic oxygen, the increase in hydrospheric and atmospheric oxygen following the filling of the major oxygen sinks depended on the sedimentary segregation of an equivalent amount of carbon. To show why, I shall simplify the photosynthetic equation to read $CO_2 + H_2O \rightleftharpoons (CH_2O)_n + O_2$. In order to create and maintain the earth's oxygen-rich atmosphere the C in the CH_2O (or glucose made from it) must be buried in the sedimentary column faster than the O_2 is consumed by recombination or by the oxidation of previously buried carbon and new reduced volcanic gases, a cycle that today takes only three million years. In other words, atmospheric oxygen is preponderantly the result of a lag in the geochemical cycling of the products of photosynthesis.

Since human beings, like most other eukaryotes, depend for their biological energy on respiratory metabolism involving the oxidation of pyruvate derived from glucose, they are likely to think of molecular oxygen as being essential to life. On the contrary, oxygen is poisonous to all forms of life in the ab-

sence of the enzymes needed to reduce destructive by-products of oxidative metabolism such as hydrogen peroxide and superoxide. Nature has also had to be clever in shielding from oxygen nuclei and other critical sites in the living cell and in designing oxidative pathways that work by the removal of hydrogen rather than by the addition of oxygen.

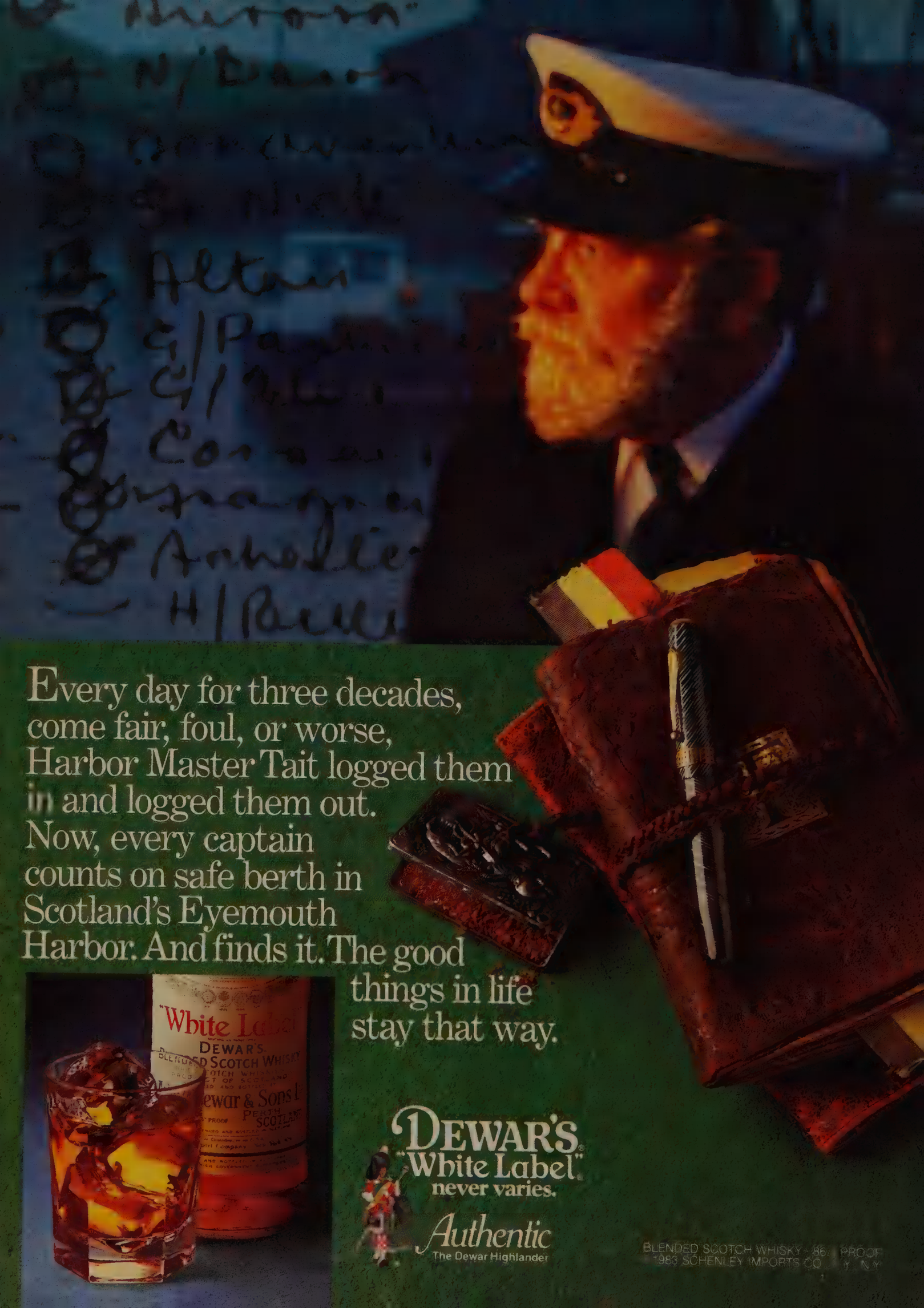
In fact, what oxygen is essential to is not the life process itself but a high level of metabolic energy. The need of eukaryotic organisms for oxygen is entirely for the production of the basic energy-transferring molecule ATP. If nonoxidative processes yielded as much ATP as oxidative processes did, there would be no demand for metabolic oxygen. It is equally certain that prebiotic chemical evolution leading to an initial stock of organic macromolecules could not have taken place in the presence of free oxygen in any form. The reactions would not go and their products would not survive. Nor could early life have survived in the presence of free oxygen before enzymic defenses against it evolved, except under vanishingly low and transient oxygen concentrations.

These indirect but well-founded biochemical conclusions are reinforced by geochemical and paleomicrobiological evidence for an initially oxygen-free earth. The record shows that the few believable fossils known before about two billion years ago were spheroidal forms and filamentous chains of cells so small and of such simplicity that they imply a prokaryotic nature and therefore a limited tolerance to oxygen. With the presence of such ample oxygen sinks as sulfide gases, sulfide minerals, ferrous iron and reduced gases from the extensive Archean volcanic belts of 2.5 billion years ago and more, these tiny organisms were under no selective pressure to acquire defenses against oxygen. Indeed, no such pressure was to arise until after the evolution of photosynthesis by blue-green algae, with its release of molecular oxygen, and after the final saturation of the major oxygen sinks.

The oldest paleontological evidence that biological oxygen was beginning to accumulate in a previously anoxic hydrosphere and thus also to escape into the atmosphere comes from the Gunflint Iron Formation of roughly two billion years ago. The filamentous microbial species *Gunflintia minuta* shows occasional enlarged cells that strikingly resemble the thick-walled cells found at intervals along the filaments of such living blue-green algae as *Nostoc*. Among such species the thick-walled cells, called heterocysts, shield the enzymes required for nitrogen fixation from the oxygen that would otherwise destroy them. Such cells lack photosynthetic pigments and neither produce nor tolerate oxygen. The similarity



BIOSPHERIC HISTORY over a span of nearly two billion years is preserved in these strata, which rise 1,340 meters from the desert floor. At the bottom are sandstone red beds that were formed in Upper Proterozoic times, 1.1 billion years ago. Just below the peak are other red beds, formed in Upper Paleozoic times, from 300 to 270 million years ago. At the crest of the peak is a formation of marine sediments that accumulated when the area was covered by sea at the end of the Paleozoic. The peak, Comanche Point, is a landmark of the Grand Canyon.


Every day for three decades,
come fair, foul, or worse,
Harbor Master Tait logged them
in and logged them out.
Now, every captain
counts on safe berth in
Scotland's Eyemouth
Harbor. And finds it. The good
things in life
stay that way.
DEWAR'S
"White Label"
never varies.
Authentic
The Dewar Highlander
BLENDED SCOTCH WHISKY · 86 PROOF
1983 SCHENLEY IMPORTS CO. N.Y.

is too close to be fortuitous. The Gunflint heterocysts imply that two billion years ago oxygen in the atmosphere had reached levels in excess of the level that now inactivates the nitrogenase enzymes. They further imply that at that time, if not earlier, there existed something very much like living nostocacean algae. Such organisms were presumably capable of splitting the water molecule as a source of the oxidative energy needed to add the citric acid cycle to an entirely fermentative metabolism.

How might the first blue-green algae, or their predecessors, have acquired such a characteristic? How might they have limited the free oxygen generated as a by-product of this activity to levels that could be managed by primitive oxygen-mediating enzyme systems? Answers to these questions are suggested by representatives of the half-dozen genera of modern blue-green algae that can switch from aerobic metabolism to anaerobic. These species prosper only in the presence of hydrogen sulfide, a reducing gas that keeps the pressure of ambient oxygen low. They have a weak tolerance for oxygen and are able to switch from water to hydrogen sulfide as a source of energetic electrons for biochemical reactions. The product of such anaerobic photosynthesis is two sulfur atoms rather than one molecule of oxygen; the sulfur atoms then become available for conversion into sulfate ion by sulfur bacteria.

Some mutant photosynthetic sulfur bacterium may have been the first to acquire reaction-energizing electrons by splitting water molecules instead of hydrogen sulfide, while at the same time retaining its ability to use hydrogen sulfide as an alternative energy source. The success of such a mutant would have been assured when its ability both to generate and to tolerate the by-product, oxygen, left it adaptively superior to microbial competitors that were exclusively anaerobic. One can imagine further mutations leading to enzymatic protection against higher oxygen concentrations while retaining (but eventually losing) access to hydrogen sulfide as an emergency energy source.

Organisms of this kind could have been responsible for unusual rock formations that were widely laid down in Archean and early Proterozoic times up to about two billion years ago. These were the finely laminated siliceous banded iron formations. Under nearly oxygen-free conditions soluble ferrous iron could have been episodically dispersed over large areas, functioning as an oxygen buffer and stimulating the growth of oxygen-producing photosynthetic microbes of limited oxygen tolerance, as hydrogen sulfide does today. Such proto-blue-green algae, in turn, could have supplied the oxygen necessary for the episodic precipitation of ferric and ferro-ferric oxides (hematite and magnetite) to make the microlaminated siliceous banded iron formations. The episodicity observed may reflect seasonal microbial blooms or episodic upwelling of ferrous iron from anaerobic basins or both. (This is not, however, to argue that *all* iron formations are so produced.) As this kind of chemical balance drew to an end the level of oxygen in the hydrosphere would have increased and so would the leakage of oxygen into the atmosphere.

The progression from a mainly anoxic hydrosphere and atmosphere to weakly oxic ones about two billion years ago is supported by two other lines of geochemical evidence. The first is the widespread presence in Africa and the Americas of the easily oxidized mineral uraninite in river sands older than about 2.3 billion years. Such extensive accumulations of so readily oxidized a mineral in stream deposits would have been improbable beneath a substantially oxic atmosphere. The other line of evidence is the near-limitation of the banded iron formations to rocks older than about two billion years. Among rocks younger than two billion years are the oldest conspicuous "red beds": sands mostly of continental origin colored by ferric oxide. Thus the banded iron formations imply a generally anoxic hydrosphere with episodes of oxidation; the red beds imply an oxidative atmosphere (and hydrosphere).

What might the level of atmospheric oxygen have been two billion years ago? Several considerations suggest that it was about 1 percent of the present level. Above that level enough ozone can form to block the harsher ultraviolet radiation of the sun. Below that level optional anaerobes can switch from aerobic metabolism to anaerobic. The apparent presence of eukaryotes by about 1.4 billion years ago implies that a level of oxygen 1 percent of the present level had been attained earlier. Uraninite in riverine sandstones up to 2.3 billion years ago indicates that at the time such a level had not yet been reached. Accordingly a continuing atmospheric level of oxygen above 1 percent is approximately bracketed between those numbers. The general transition from banded iron formations to red beds about two billion years ago suggests the passing of some threshold level of oxygen about then. Was that the first attainment on a sustained basis of 1 percent of the present atmospheric level of oxygen? Quite likely.

If the most important event in the evolution of the biosphere was the first appearance of life, it is run a close second and third by the appearance of chlorophyll *a* as the mediator of oxygen-producing photosynthesis sometime before two billion years ago and the appearance of the eukaryotic cell, with its characteristic mitotic cell division, between two and 1.4 billion years ago. Whereas in the prokaryotic cell the DNA forms a single long chromosome that is folded irregularly throughout the cell, in the eukaryotic cell the chromosomes are numerous, rodlike and shielded within a well-defined, membrane-bounded nucleus. In the process of mitosis the chromosomes cluster into a central spindle and split into pairs before each cell division. Mitosis depends on the contractile properties of the protein actomyosin, which cannot form in the absence of oxygen. The more advanced steps in the synthesis of sterols, fatty acids and the fibrous protein collagen, leading to muscles and metazoans, are also dependent on a sufficient level of oxygen.

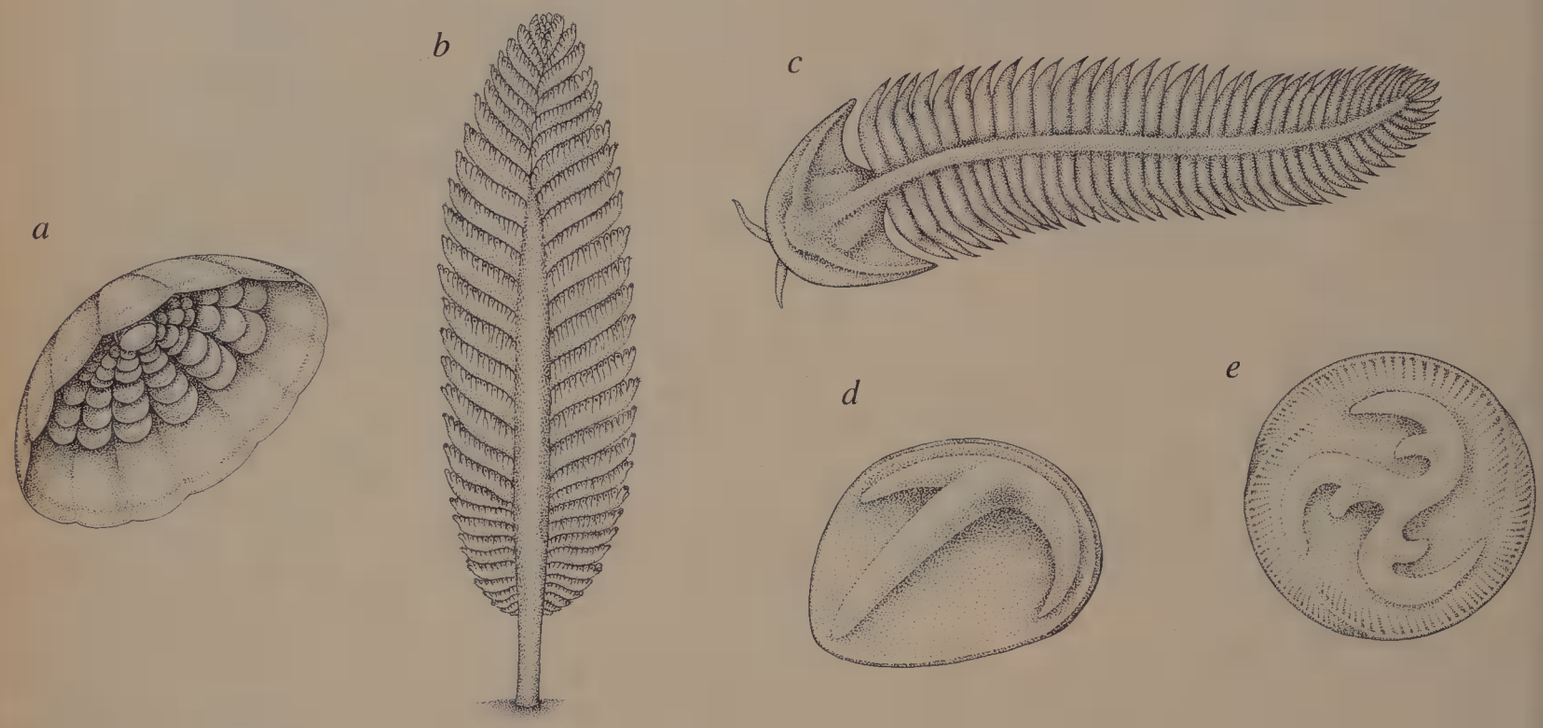
Still, it remains unsettled exactly how either eukaryotic cells or the metazoans arose. Part of the story of the origin of the eukaryotic cell surely involves endosymbiosis: the engulfing of one organism by another, giving rise to cell organelles such as the mitochondrion and the chloroplast. Yet unknown processes are just as surely involved in the origin of the distinctive eukaryotic nucleus, mitosis and meiosis: the reductive division of chromosomes for the purpose of reproduction. The gulf between the prokaryote and the eukaryote is comfortably bridged, however, by the identity in all organisms of the genetic code, the universality of ATP as the energy-transferring molecule and the consistency of the amino acid composition of proteins in all forms of life. Such universals make sense only if they reflect a common ancestry.

What is the fossil evidence for the appearance of eukaryotes between two and 1.4 billion years ago? For one thing eukaryotic cells are generally larger than prokaryotic cells. As V. V. Timofeev and other Russian paleomicrobiologists have long noted, the average cell diameter of fossil microorganisms increases substantially in rocks of younger Proterozoic age. James W. Schopf of the University of California at Los Angeles has compiled data implying that the influx of larger cells took place about 1.4 billion years ago and involved a shift from sizes of generally less than 10 micrometers to commonly more than 20 micrometers. G. R. Licari and I have observed cell diameters up to 60 micrometers among microfossils of eastern California that may be 1.3 billion years old. Such dimensions imply a change then or earlier from an entirely prokaryotic microflora to a partly eukaryotic one. The geologic record for the interval between two and 1.4 billion years ago, however, is still sufficiently incomplete to leave open the possibility that older eukaryotes may yet come to light. It seems that oxygen levels would have favored their appearance at

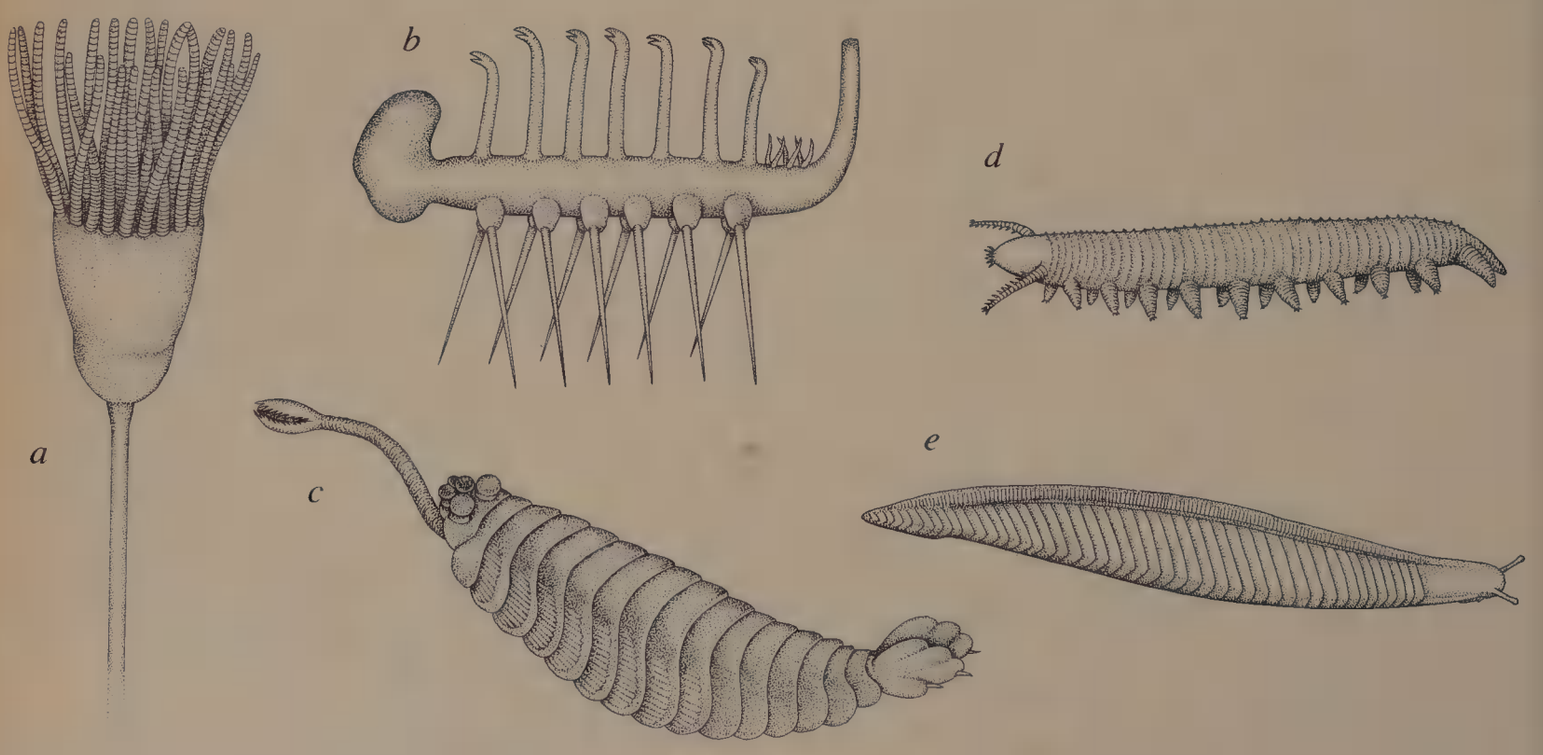
any time from about two billion years ago onward.

All in all there is little reasonable doubt that at some time between two and 1.4 billion years ago eukaryotic microorganisms, probably including green and red algae, became established. This opened the way to the biological precipitation of silica, which only eukaryotes can carry out (and which is probably why the silica in the banded iron formations seems to have been precipitated not biologically but chemically). It also satisfied one of the two essential preconditions for the evolution of the metazoans: the existence of the eukaryotic cell. The other precondition to be met, as Kenneth M. Towe of the Smithsonian Institution has pointed out, was a level of free oxygen high enough for the manufacture of such products as collagen.

After about 1.4 billion years ago the previously skimpy record of preserved cellular remains gets much better. It is as if evolution were speeding up. Yet it is another 700 million years before there is



PRECAMBRIAN ANIMALS, once best known from the Ediacara Hills of South Australia, have now been found on five continents. The most abundant animals resemble jellyfish (*a*) or other modern coelenterates such as the coral-like ones known as sea pens (*b*). Others, such as *Spriggina* (*c*) resemble naked arthropods and annelid worms. Still others, such as *Parvancorina* (*d*) and *Tribrachidium* (*e*), resemble no other known animal. All probably took their supply of oxygen from the surrounding water by absorption through their epithelia.



CAMBRIAN ANIMALS, far more abundant than the preceding Ediacarian ones and commonly protected by an exoskeleton, also included soft-bodied species that are best preserved in a Middle Cambrian formation at Mount Robson in Canada. By this time representatives of three new and now extinct phyla had appeared: the immobile *Dinomischus* (■), a stilt-legged scavenger, *Hallucigenia* (*b*), and the predaceous *Opabinia* (*c*). A fourth newcomer, *Aysheaia* (*d*), resembles the living onychophoran *Peripatus*. Probably the most advanced of the group, *Pikaia* (*e*), is interpreted as the sole representative of the chordates, the phylum that gave rise to all vertebrate life.

evidence of the oldest fossil organisms that all agree to be at once metazoan, indigenous to the sedimentary deposits in which they are found and deposited concurrently with those sediments. These organisms are the soft-bodied marine invertebrates of the Ediacarian System, which can be placed in the earliest part of the Paleozoic era and the Phanerozoic eon.

The Ediacarian fauna was originally known only from the Ediacara Hills of South Australia, but representatives of an even larger assemblage of fossils of this age have now been found in some two dozen different areas on five continents, entombed in sedimentary rocks deposited between 670 and 550 million years ago. Their appearance followed a nearly global succession of glacial deposits in later Proterozoic time. From such beginnings and from contemporaneous algal life there eventually arose the familiar biosphere of today.

The dominant organisms of the Ediacarian System, accounting for nearly 70 percent of the total, are coelenterates (more precisely Cnidaria), and of the coelenterates three-fourths are either jellyfishlike floaters or colonial forms reminiscent of the modern siphonophore *Velella,* drifters of the open sea and stranders on many a beach. The remaining coelenterates are colonial types that were attached to the sea floor like the modern sea pens. The noncoelenterates of the Ediacarian fauna include marine worms resembling modern polychaetes, some unusual arthropodlike animals that lack a carapace and a curious three-rayed discoidal animal that suggests a tiny naked starfish.

Some of these early drifters and inhabitants of the shallow sea floor left surface tracks and body imprints as records of the past, but none seems to have burrowed vertically into the bottom. Some were also quite large: jellyfish up to a meter in diameter and sea pens more than a meter long. A sheetlike marine worm called *Dickinsonia* grew to a length of nearly a meter but was less than three millimeters thick.

Although the Ediacarian animals are primitive, they are scarcely the kind of near-microscopic fauna that many paleontologists have expected to find at the base of the metazoan lineage. Still, in retrospect these strange animals are not unlike what one might expect if the reaching of some critical level of dissolved oxygen had triggered their evolution. As Rudolf A. Raff and Elizabeth C. Raff of Indiana University have demonstrated, jellyfish and similar coelenterates can get their oxygen through surface absorption at concentrations equivalent to only about 7 percent of the present atmospheric oxygen level. Under such conditions a thin, metabolically active surface combined with a large oxygen-collecting area would be advantageous. Thus it should not be surprising that jellyfish, thin-bodied worms and naked arthropods should predominate in the oldest known animal life. There is evidence that *Dickinsonia* had a gut and a weakly muscular body, implying that it also had an internal oxygen-collecting system. The Australian paleontologist Bruce Runnegar has calculated that even this form could have acquired enough oxygen at dissolved-oxygen pressures as low as from 6 to 10 percent of those near the sea surface today.

The fact that eukaryotic cells had apparently been around for 700 million years or more before the oldest known metazoans came along suggests that the triggering event for metazoan evolution was the cumulative buildup of oxygen in the hydrosphere and the atmosphere to somewhere near 7 percent of its present level. Quite possibly such a triggering was abetted by the ecological stress and geographic isolation arising from the drift of lithospheric plates, which at the time was rapid, and the related climatic changes. The first animals with phosphatic and calcareous external skeletons are found among the oldest Cambrian fossils and in rare instances among the youngest Ediacarian ones. Such impervious coverings would exclude the surface absorption of oxygen. By then gills and circulatory systems of some kind had presumably begun to function, suggesting an oxygen level perhaps closer to 10 percent of the present one. The race toward the more efficient utilization of biological energy was on.

Feedback effects related to the photosynthetic buildup of oxygen were far-reaching. In addition to its stimulation of eukaryotic and metazoan evolution, oxygen increased at the expense of carbon dioxide in both the hydrosphere and the atmosphere. The higher early carbon dioxide pressures probably account for the prevalence of dolomite over limestone among marine sediments during much of pre-Phanerozoic time. Nitrogen, which is the dominant atmospheric gas today, probably became so only gradually. Nevertheless, it was presumably always important. In the distant past it was a vital biospheric nutrient. It remains one today and is perhaps equally important as a nearly inert dilutant for corrosive oxygen.

The rest of biospheric history reflects the response of eukaryotic evolution to plate tectonics, climate and ecological challenge on the evolving Phanerozoic earth. It is a matter of detail, much of it beautiful and rich in significance. The Phanerozoic biosphere, descended from an almost entirely microbial one, has had a long history of ever more elaborate diversification that is still in progress. Where will it end? End it must, if only when the sun becomes a red-giant star four or five billion years from now.



There's a lot worth saving in this country.

Today more Americans who value the best of yesterday are saving and using old buildings, waterfront areas and even neighborhoods.

Preservation saves energy, materials and the artistry of these quality structures.

Help preserve what's worth saving in your community. Contact the National Trust, P.O. Box 2800, Washington, D.C. 20013.

National Trust for Historic Preservation
Preservation builds the nation

AUDIO-FORUM® offers the best in self-instructional foreign language courses using audio cassettes — featuring those used to train U.S. State Dept. personnel in Spanish, French, German, Portuguese, Japanese, Greek, Hebrew, Arabic, Chinese, Italian, and more.

Learn a foreign language on your own! Free Catalog.

Call **(203) 453-9794**, or fill out and send this ad to –

Audio-Forum
Suite 35L, On-the-Green
Guilford, CT 06437

Name

Address

City

State/Zip

I am particularly interested in (check choice):

☐ Spanish ☐ French ☐ German ☐ Polish
☐ Greek ☐ Russian ☐ Hausa ☐ Czech
☐ Bulgarian ☐ Turkish ☐ Vietnamese
☐ Other

THE AMATEUR SCIENTIST

Caustics: mathematical curves generated by light shined through rippled plastic

by Jearl Walker

Caustics are patterns of bright points and lines that form when light reaches a surface by refraction or reflection. You can see them on a tablecloth in the light refracted by a glass of white wine. They can also form on the bottom of a swimming pool when sunlight is refracted by the waves at the surface. My favorite example appears when a laser beam is directed through a piece of irregularly rippled plastic and illuminates a screen with a beautiful array of bright lines. Even a slight motion of the plastic across the beam sends the patterns of light dancing into variegated new designs.

The caustic patterns can be simple or complex. Different materials lead to different designs. Is the number of possible patterns endless, or is there some way to classify them in groups of basic designs? Until recently I supposed the number of patterns was infinite. New developments in optics, however, have revealed a tidy way of analyzing the patterns. It turns out that there are only a few basic designs, which are called the elementary catastrophes.

This classification is based on catastrophe theory, a mathematical analysis originated in the 1970's by René Thom of the Institut des Hautes Études Scientifiques at Bures-sur-Yvette in France. The application of the theory to optical caustics was the work of Michael V. Berry of the University of Bristol. I have based my study on this work.

To examine caustic patterns I directed the beam from a helium-neon laser through a layer of transparent plastic with a rippled surface. The plastic, which came from the cover over a fixture of fluorescent light bulbs, had what appeared to be a random arrangement of smooth hills and valleys rather than a regular pattern. At some distance from the plastic was a screen on which the caustics appeared.

By moving the plastic through the beam I could rapidly sample many caustic patterns. Some were simple curved lines. Others were quite complex with overlapping bright lines. Interference patterns also appeared with the caustics, indicating that the light waves interfered destructively and constructively at the screen. I concentrated only on the caustics.

The caustic patterns can be separated into several basic units. The commonest type is a smoothly curved line that results from what is called a fold catastrophe. In addition the pattern might have a bright point, a cusp, a swallowtail, a triangle, a butterfly or a corner. These basic units follow from the several elementary catastrophes arising when the beam from the laser refracts through the rippled plastic.

I also obtained caustic patterns from a glass slide coated with an uneven layer of plastic glue. Laser light is advantageous in these experiments because its rays normally spread only slightly. Sunlight, spreading much more, tends to obscure some of the basic caustic designs. Another source of light might serve if it is far enough from the refracting material to appear to be a point source of light. Only then is the spread in the rays from the source sufficiently small so that the refracting material can yield clear caustic designs.

An intriguing caustic pattern can be seen readily in sunlight. Put a drop of water on a glass slide and hold it just above a flat surface. On the surface a pattern appears that almost always has cusp caustics around the perimeter.

To explore catastrophe theory in optics I shall consider the caustics that can be produced by a plane wave of light passing through a transparent material such as a layer of plastic. Before the light reaches the plastic it is traveling in the positive direction of the z axis in the top illustration on page 196. The notion of a wave surface aids in tracking the progress of the light wave. This surface is an imaginary one on which all parts of the wave are in phase. For example, if light is considered to be a wave of crests and troughs, at some instant only crests pass through the imaginary surface. A short time afterward only troughs pass through.

The shape of the wave surface is given mathematically by a function f that is the distance between any point on the wave surface and an underlying x-y plane that serves as a reference plane. When the light is a plane wave, f is simple because the wave surface is flat and parallel to the x-y plane. Thus all points on the wave surface are at the same distance from the x-y plane, and f is merely a constant.

The notion of a light ray also helps in visualizing the progress of a wave. A ray is a vector pointing in the direction of travel of the light. When a ray is added to a section of the wave surface, it is drawn perpendicular to that section. If the light wave is planar, the light rays are easy to draw because they are all parallel.

When the light wave passes through a layer of rippled plastic, the wave surface and the rays are no longer as easy to draw. I shall describe several possible results since generally the shape of the plastic surface is not known. For simplicity I assume two limitations on the shape. First, the surface must be smoothly rippled, with no sharp ridges or intentional patterns. A sharp edge complicates the analysis, and an intentional pattern might dominate the pattern of light cast on the screen. Second, the ripples on the surface should be larger than the wavelength of visible light so that the light pattern is not due merely to wave interference.

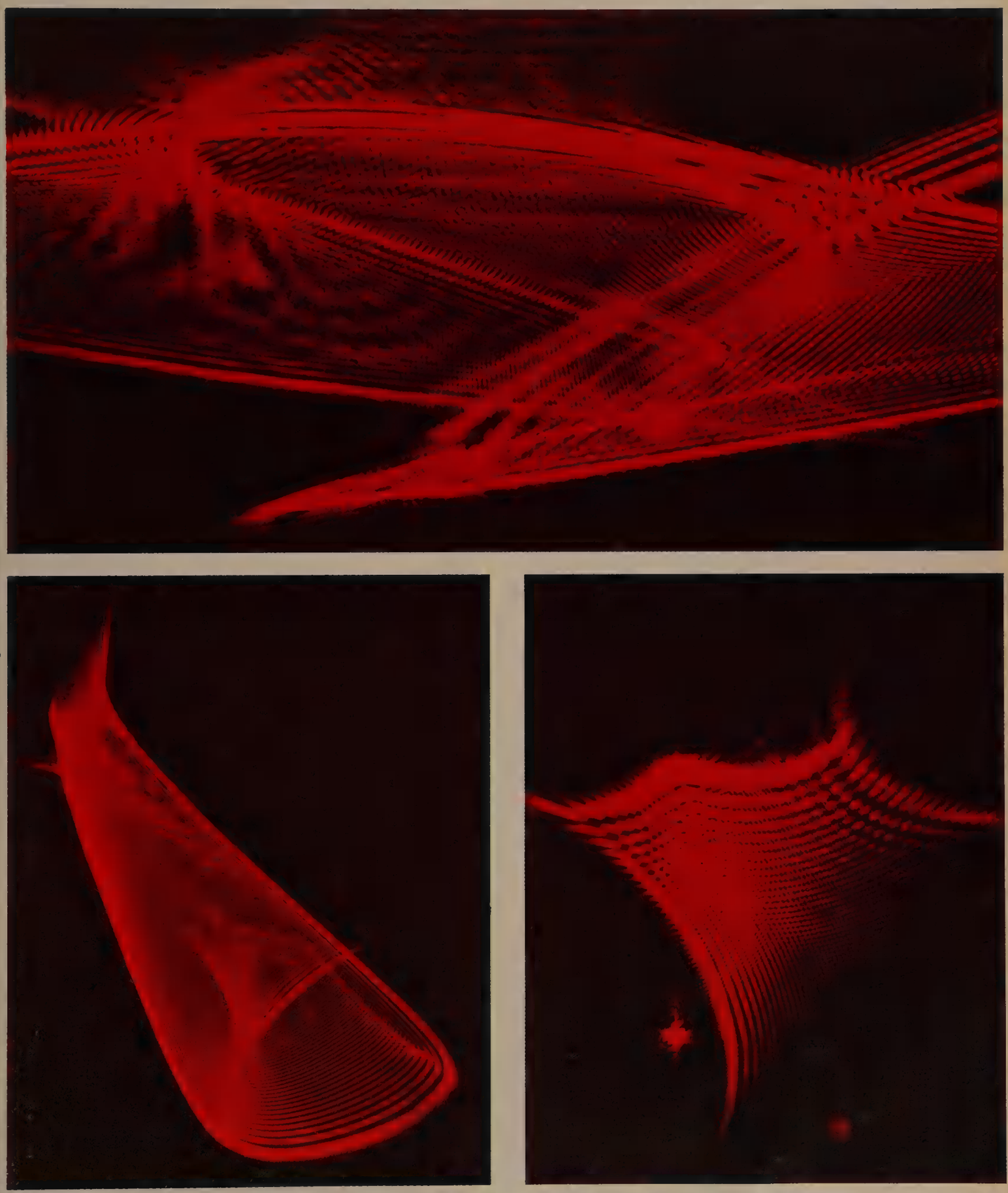
When the light wave passes through the plastic layer, it is refracted in many directions. Hence its wave surface is no longer flat and the light rays are no longer parallel to the z axis. The value for the function describing the height of a point on the wave surface is no longer a constant. A point on a hill on the wave surface is far from the x-y plane and a point in a valley is closer.

When rays are added to the picture, they must again be perpendicular to the wave surface wherever they are drawn. Since that surface is no longer flat, the rays point in many directions. The part of the light wave passing through a given section of the wave surface travels in the direction of the ray assigned to that section. The travel of the entire light wave is therefore harder to follow once the light passes through the plastic.

Much farther along the z axis lies the screen. That is where you see the results of the distortion of the light wave by the plastic. If the plastic were flat, the screen would be evenly illuminated. With rippled plastic much of the screen may be partially illuminated, but in some places the bright caustics will appear. Each point on a caustic is formed because the plastic layer bunches many rays onto that region of the screen.

My objective was to classify the kinds of pattern that can appear on the screen without specifying any details about the surface structure of the plastic layer. The task is to figure out what kinds of pattern are possible on the screen by studying what kinds of shape a wave surface can have. Catastrophe theory predicts that the wave surface can have only a limited number of distinguishable shapes. Thus only a limited number of basic caustic patterns can appear on the screen. Since any complicated pattern of caustics can be broken down into these basic patterns, one can immediately describe the shape of the wave surface even though the details about the surface of the plastic are not known.

To determine the possible shapes of the caustics one must examine the curvature of the wave surface. The bottom



An array of laser-beam caustics

CAN A COMPUTER MAKE YOU CRY?

Right now, no one knows. This is partly because many would consider the very idea frivolous. But it's also because whoever successfully answers this question must first have answered several others.

Why do we cry? Why do we laugh, or love, or smile? What are the touchstones of our emotions?

Until now, the people who asked such questions tended not to be the same people who ran software companies. Instead, they were writers, filmmakers, painters, musicians. They were, in the traditional sense, artists.

We're about to change that tradition. The name of our company is Electronic Arts.

SOFTWARE WORTHY OF THE MINDS THAT USE IT. We are a new association of electronic artists united by a common goal — to fulfill the enormous potential of the personal computer.

In the short term, this means transcending its present use as a facilitator of unimaginative tasks and a medium for blasting aliens. In the long term, however, we can expect a great deal more.

These are wondrous machines we have created, and in them can be seen a bit of their makers. It is as if we had invested them with the image of our minds. And through them, we are learning more and more about ourselves.

We learn, for instance, that we are more entertained by the involvement of our imaginations than by passive viewing and listening. We learn that we are better taught by experience than by memorization. And we learn that the traditional distinctions — the ones that are made between art and entertainment and education — don't always apply.

TOWARD A LANGUAGE OF DREAMS. In short, we are finding that the computer can be more than just a processor of data.

It is a communications medium: an interactive tool that can bring people's thoughts and feelings closer together, perhaps closer than ever before. And while fifty years from now, its creation may seem no more important than the advent of motion pictures or television, there is a chance it will mean something more.

Something along the lines of a universal language of ideas and emotions. Something like a smile.

The first publications of Electronic Arts are now available. We suspect you'll be hearing a lot about them. Some of them are games like you've never seen before, that get more out of your computer than other games ever have. Others are harder to categorize — and we like that.

WATCH US. We're providing a special environment for talented, independent software artists. It's a supportive environment, in which big ideas are given room to grow. And some of America's most respected software artists are beginning to take notice.

We think our current work reflects this very special commitment. And though we are few in number today and apart from the mainstream of the mass software marketplace, we are confident that both time and vision are on our side.

Join us.
We see farther.

ELECTRONIC ARTS™





TO LEARN MORE about our growing number of titles—and to receive a free poster of the artists pictured here—stop by your favorite computer store or software center. If you need help finding the Electronic Arts dealer nearest you, write us at 2755 Campus Drive, San Mateo, California 94403 or call (415) 571-7171.

SOFTWARE ARTISTS? "I'm not so sure there *are* any software artists yet," says Bill Budge. "We've got to earn that title." Pictured here are a few people who have come as close to earning it as anyone we know.

That's Mr. Budge himself, creator of PINBALL CONSTRUCTION SET, at the upper right. To his left are Anne Westfall and Jon Freeman who, along with their colleagues at Free Fall Associates, created ARCHON and MURDER ON THE ZINDERNEUF.

Left of them is Dan Bunten of Ozark Softscape, the firm that wrote M.U.L.E. To Dan's left are Mike Abbot (top) and Matt Alexander (bottom), authors of HARD HAT MACK. In the center is John Field, creator of AXIS ASSASSIN and THE LAST GLADIATOR. David Maynard, lower right, is the man responsible for WORMS?

When you see what they've accomplished, we think you'll agree with us that they can call themselves whatever they want.

In the flash of one second, Delta-10 can print the alphabet six times.

At 160 cps, 100% duty cycle, Delta-10 is built for speed. It's the dot matrix printer that quickly strikes up an intense working relationship with your computer.

There's no nonsense. Delta-10 performs with serial and parallel interfaces as standard. The 8K buffer gives Delta-10 the power to instantaneously store and print data as it frees your computer to continue its job. Turn Delta on and it can print three crisp copies simultaneously at the continuous speed of 160 cps.

And Delta-10's performance isn't only measured in swiftness. There's the diversity of its character fonts, its true descender matrix, and its underlining ability.

Star's new Delta-10. Designed for red-hot efficiency. And isn't that what you need?

star micronics • inc

THE POWER BEHIND THE PRINTED WORD.

Computer Peripherals Division
P.O. Box 612186, Dallas/Ft. Worth Airport, TX 75261



illustration on the next page shows a cross-sectional slice of a hill on the surface. The x axis of the underlying x-y plane is included. I am interested in how the slice curves with respect to the x axis. Two regions of the slice are curved and one section (the inflection on the side of the hill) is not. Rays from the curved sections are spread over the screen with no bunching. Rays from the straight section are bunched onto one spot on the screen and so contribute to a caustic.

The figure is for only one cross-sectional slice of the wave surface. Next to that slice another one can be depicted. Again part of the underlying x-y plane is to be drawn below the slice. For convenience I call this direction x too even though it probably differs from the direction in the first slice.

Part of the hillside of the second slice lacks curvature. Perhaps this region is a bit higher or lower on the hillside than it was before, but such a detail is not important just yet. The point is that one can always take a slice of any part of the wave surface and examine it for curvature. If part of the slice lacks curvature, it can contribute to a caustic to the screen.

Suppose all the uncurved places on the wave surface are discovered. On the underlying x-y plane imagine a line L that is just below those places. The shape of L determines the shape of the caustics on the screen. Usually the line creates a smoothly curved line of light on the screen, but it can create several other caustic patterns: the other elementary catastrophes.

A graphic analysis greatly aids one's use of catastrophe theory. Reconsider the illustration of a cross-sectional slice through a hill on the wave surface. Rays from the hillside end up at various places on the screen. The axis a shows where they shine. (The illustration is misleading because the screen is necessarily shown close to the wave surface.)

Rays from the bottom of the hill point directly toward the screen. As one ascends the hill the rays began to point more to the left. The extreme is reached in the interesting region that has no curvature. Further ascent of the hill takes one through rays pointing less to the left. Finally, on the hilltop the rays are again pointing directly toward the screen.

Since the screen is distant from the wave surface, the rays both from the flat bottom of the hill and from the hilltop end up along the same region of the screen. The rays that are deflected to the left side of the screen are those from the side of the hill. Points along the wave surface (from the bottom of the hill to the top) are connected by the rays to points along the screen.

The connection is best expressed in terms of the values of x (on the axis below the wave surface) and a (on the axis across the screen). The upper illustration at the right on page 197 shows the relation of the values. For large values of a two regions on the wave surface contribute a ray. For example, the top and bottom sections of the hill both send rays to approximately the same place. With classical optics one can calculate the intensity of the light at that point on the a axis. Since the rays are not bunched, the intensity is less than it is at a caustic.

Near the fold in the curve the value of a reaches a limit set by the rays from the side of the hill. The curve folds over for the limiting value, which means many rays arrive at that point on a from a range of places on x. Classical optics predicts the intensity to be infinitely large at the foldover point on a. This is a caustic point.

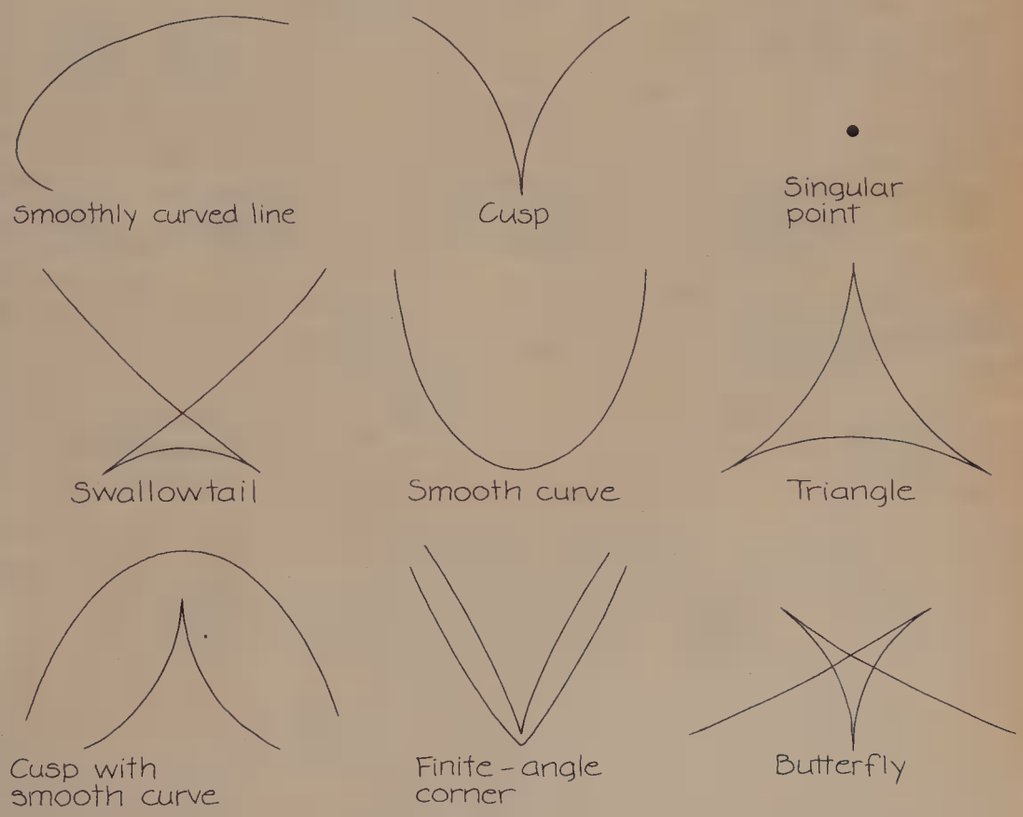
This arrangement is a fold catastrophe. The term implies that the graphic relation between points x (associated with the wave surface) and points a (along the screen) has a fold. The word catastrophe is appropriate too. Points on a away from the caustic have a simple contribution of rays, one ray from a large value of x and another from a small one. As one considers points of a nearer the fold, however, the rays abruptly bunch to produce a bright spot at the fold point on a.

The algebraic relations between values of x and a for the elementary catastrophes are usually given by means of generating functions. The functions are listed in the bottom illustration on page 197. For the fold catastrophe two variables are required. The a is called a control variable, the x a state variable. To gain the relation between x and a for the fold catastrophe its generating function is differentiated with respect to the state variable x and set to zero. The resulting equation is the one plotted in the upper illustration at the right on page 197.

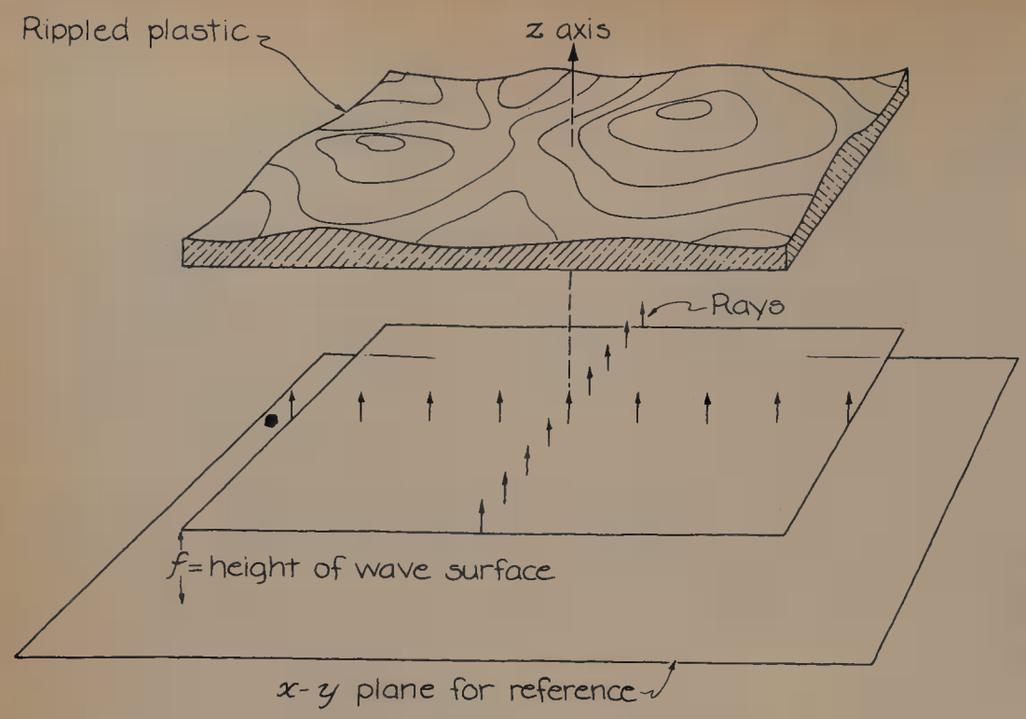
I have now demonstrated that a section on the wave surface can create a bright point on the screen. Suppose that next to this section on the wave surface other sections also contribute a bright point apiece. The string of bright points on the screen forms a smoothly curved caustic line, the commonest kind of pattern yielded by a rippled layer of plastic inserted in a laser beam. Each point on the caustic line is the result of a fold catastrophe from a section on the wave surface that lacks curvature with respect to a direction on the underlying x-y plane. The line L running below these sections on the wave surface is transformed by means of the rays into the caustic line on the screen.

A cusp catastrophe is somewhat more complex. It involves the same state variable x for the wave surface but includes a new control variable b on the screen in addition to the a already employed. In other words, in a cusp catastrophe the points x for the wave surface determine the rays in the region of a and b on the screen. The algebraic relation between x, a and b is obtained by differentiating the appropriate generating function by the state variable x and then setting the result to zero.

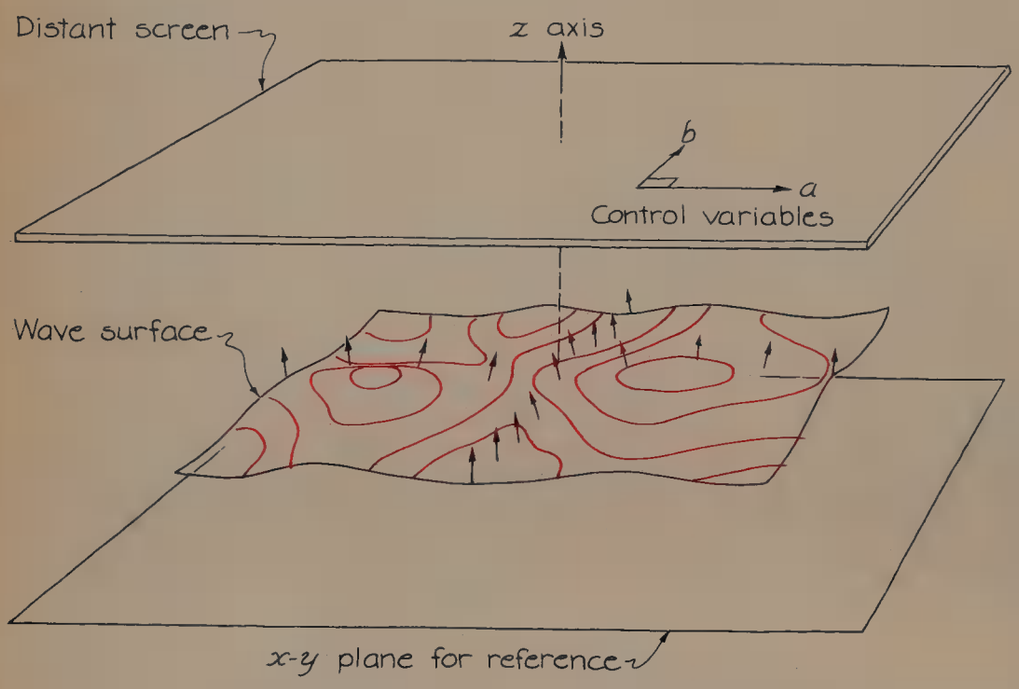
Again a graph aids one's understanding of the algebraic relation. This time



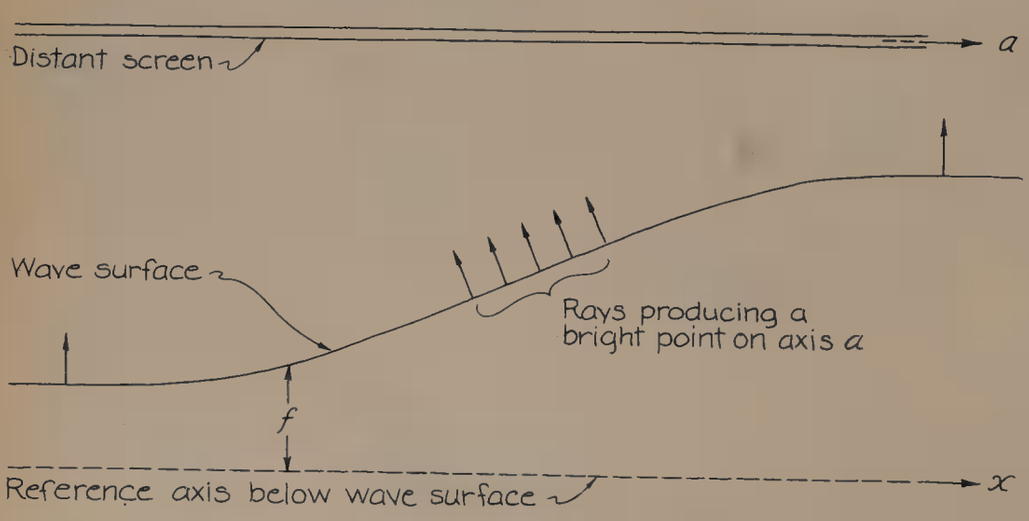
A classification of caustic patterns



The surface of a light wave before it reaches the plastic



The wave surface after it has passed through the plastic



Rays from a hillside on the wave surface

the graph is three-dimensional. As is shown in the top illustration on page 199 two dimensions of the graph are devoted to the control variables *a* and *b;* the vertical dimension is the state variable *x*.

The bottom plane of the graph is the control space and actually represents the screen. Above it is a folded sheet whose shape is set by the equation obtained from differentiating the generating function. (One should not be misled into believing this sheet is visible or tangible. It is only a mathematical relation between the points labeled *x* on the wave surface and points *a* and *b* on the screen.)

What one sees on the screen is the projection of the folded sheet onto the *a*-*b* surface of the graph. The projection is a bright cusp. Hence if a layer of rippled plastic produces the right kind of wave surface, one finds a caustic cusp on the screen instead of the smoothly curved caustic line formed from a series of simple fold catastrophes.

What kind of wave surface produces a cusp? A fold catastrophe results from a region on the wave surface that has no curvature along some direction in the *x*-*y* plane. I call that direction *x* for any slice through the wave surface. The line *L* runs through the points on the *x*-*y* plane that are below all these places on the wave surface without curvature. At each point along *L* I can construct *x* and *y* axes so that *x* is in the direction of no curvature for the part of the wave surface just above the point.

In general the *x* axis at each point is in a direction different from that of the line *L* through the point. If the surface of the plastic is appropriately shaped, however, the wave surface can have a region in which the *x* axis coincides with the direction of *L* there. The rays from this region end up forming a bright cusp on the screen.

The bottom illustration on page 199 indicates a candidate for such a region on the wave surface. The regions without curvature form a path that runs along the side of a hill, climbs the hill and then again runs along the side. Consider the shape of *L* just below the path. Also consider the direction of no curvature for points along *L*. For any point on *L* away from the climbing region the direction of no curvature is not coincident with *L*. Those points therefore contribute only simple fold catastrophes. Seen as a composite they are smoothly curved caustic lines on the screen. In the climbing regions the direction of *L* coincides with the direction of no curvature. Hence rays from that region produce the sharp point on the cusp caustic on the screen.

The graph of the cusp catastrophe illustrates the connection between points of *a* and *b* on the screen and points *x* for a path over a wave surface. As before

the connection is by means of the rays leaving the hillside for the screen. To show the connection I have numbered five sections on a path over the hillside in order to show where rays leaving those sections end up on the screen. Bear in mind again that the screen is actually much farther from the wave surface than the illustration suggests. Remember too that the cusp is likely to be much larger than the wave surface.

Rays from sections 1, 3 and 5 end up in the middle of the cusp. As one climbs the hill through section 2 the rays are deflected to the left side of the screen. Along section 2 the rays begin to bunch to form the left side of the cusp caustic. Higher on the hill the rays begin to fall inside the cusp again. Section 2 is therefore responsible for the extreme left side of the caustic pattern.

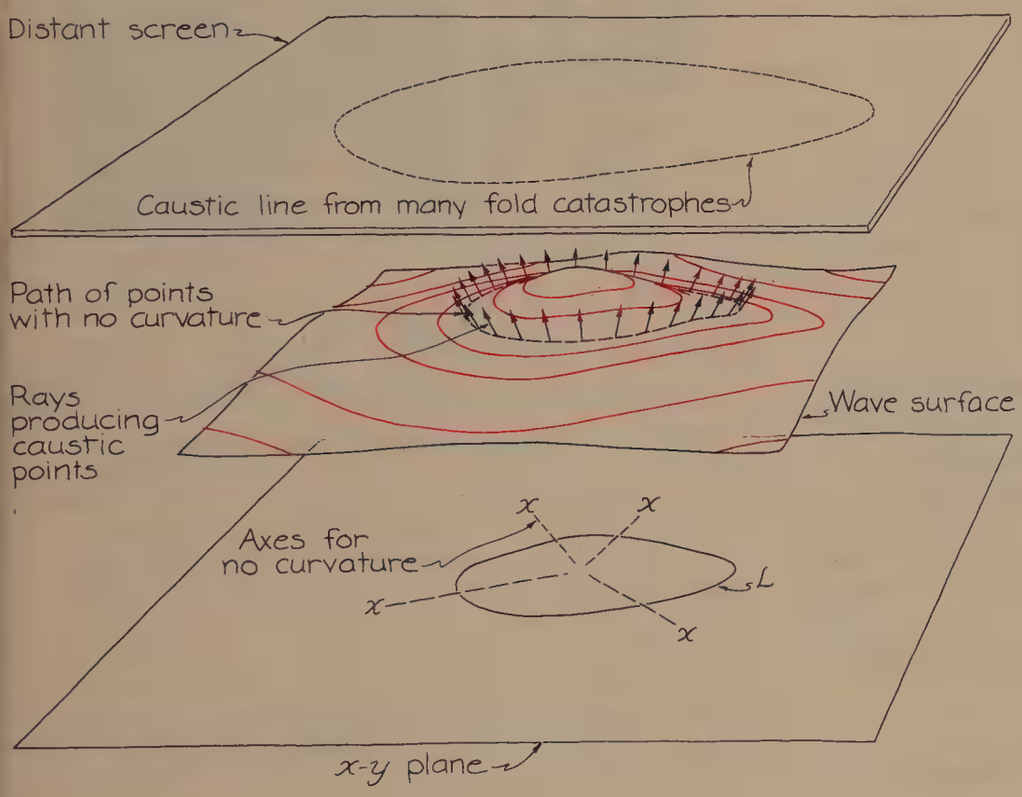
As one begins to descend the hill through section 4 the rays are sent off to the right side of the screen. The bunching of rays from near section 4 is responsible for the caustic line on the right. Rays from below section 4 on the hill fall inside the cusp.

A similar analysis can be made on the three-dimensional graph of the cusp caustic. In the illustration I have rotated the cusp so that the fold in the graph can be visualized. Again I have added labels to indicate where the rays originated from the hill on the wave surface. The edge of the cusp labeled 4 lies below the fold on the upper sheet of the graph. The other edge lies below the fold on the lower sheet. Thus the upper fold represents the connection between the right side of the hill and one edge of the cusp, whereas the lower fold is for the left side of the hill and the other edge.

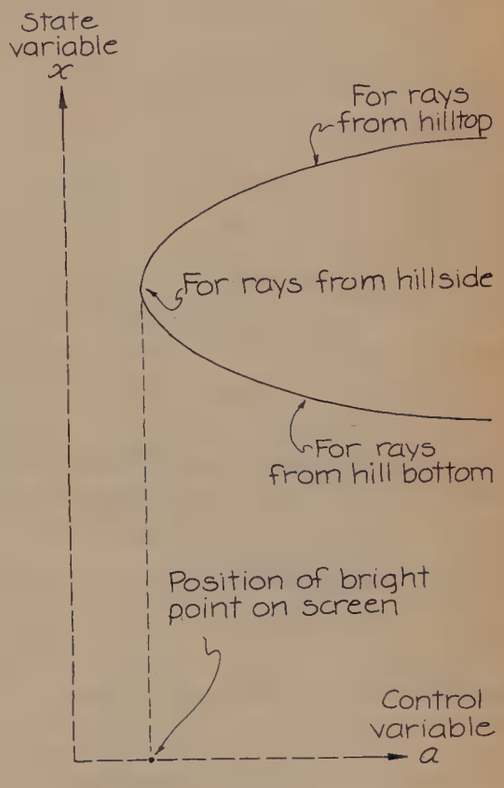
Outside the cusp region the screen is not illuminated. The sections of the folded sheet above these dark regions are extraneous and could be eliminated from the graph. Points at the center of the cusp lie below three sheets. The top sheet represents the connection for the rays leaving the hill on the right side, below section 4. The middle sheet is for the rays from the top of the hill. The bottom sheet is for the rays from the section numbered 1.

Next consider a part of the folded sheet closer to the point where the fold smoothes out. That point is called a singularity. There the graph represents rays that originate in the climbing region and fall on the cusp point on the screen. Farther back on the graph the sheet is not folded. Below this section there is no caustic.

Part of catastrophe theory deals with classifying sections of a catastrophe as either generic or nongeneric. The definitions of these terms can be demonstrated by two lines on the screen that slice



How a smoothly curved caustic line is formed



A graph of a fold catastrophe

Catastrophe	Control variables	State variables	Generating function	First derivative
Fold	a	x	$\frac{1}{3}x^3 - ax$	$x^2 - a = 0$
Cusp	a,b	x	$\frac{1}{4}x^4 - ax - \frac{1}{2}bx^2$	$x^3 - a - bx = 0$
Swallowtail	a,b,c	x	$\frac{1}{5}x^5 - ax - \frac{1}{2}bx^2 - \frac{1}{3}cx^3$	$x^4 - a - bx - cx^2 = 0$
Butterfly	a,b,c,d	x	$\frac{1}{6}x^6 - ax - \frac{1}{2}bx^2 - \frac{1}{3}cx^3 - \frac{1}{4}dx^4$	$x^5 - a - bx - cx^2 - dx^3 = 0$
Hyperbolic umbilic	a,b,c	x,y	$x^3 + y^3 + ax + by + cxy$	$3x^2 + a + cy = 0$ $3y^2 + b + cx = 0$
Elliptic umbilic	a,b,c	x,y	$x^3 - xy^2 + ax + by + cx^2 + cy^2$	$3x^2 - y^2 + a + 2cx = 0$ $-2xy + b + 2cy = 0$

The algebraic relations for six elementary catastrophes

EXPLORE THE CHALLENGES OF THE NUCLEAR AGE —

Participate in a wide-ranging, open dialogue among scientists, scholars, government officials and concerned citizens in the Bulletin of the Atomic Scientists.

"The Bulletin is unique. In my judgment it is the definitive magazine on science and world affairs."

--Hans Bethe,
Cornell University
Nobel Laureate



Read why arms expert Richard Garwin says many American defense contracts are based on hidden agendas or misinformation. Join the debate on the arms race between Edward Teller and Frank von Hippel. Get an insider's view of the war in Afghanistan in an article illustrated with photos taken by an American who wandered the countryside posing as an Afghani. Follow a discussion of government secrecy and scientific research. Look inside the politics, culture, religion and military establishment of the Soviet Union in a special **Bulletin** series. Be privy to the latest information on arms negotiations and military policy.

If the subject is about science and world affairs, it's sure to be discussed in a lively and informative manner in the **Bulletin**. Try a free issue and explore the crucial challenges of the nuclear age with the **Bulletin** . . .



1947 1963 1983

The Bulletin's Clock is the internationally recognized symbol of how close the world is to nuclear disaster. The clock represents not only peril, but hope that solutions will be found to create a secure and peaceful world.

Special offer for Scientific American readers: Receive the next issue of the Bulletin *free* and with no obligation.

1) We will send you our next issue *free* when you mail the tear-out card or call collect (312) 363-5225.

2) After reading your free issue, if you decide to continue as a subscriber, you *save* $10 off the single issue price, or $4.50 off our regular subscription price, with this special offer for *Scientific American* readers.

3) You also receive, as a **free bonus**, the **Bulletin's** historic edition, "40 Years of the Nuclear Age." This 80-page issue marking the 40th anniversary of the first nuclear chain reaction will be mailed promptly upon receipt of your payment.

Right now you can enjoy a full year (10 issues) of the **Bulletin of the Atomic Scientists** for only $15 — just $1.50 per copy! And you have our complete *guarantee*: you may cancel at any time and receive a refund on the unmailed copies.

Use the tear-out card to send for your free issue today. Remember, there is no obligation to subscribe.

Bulletin
of the Atomic Scientists **5801 South Kenwood Avenue, Chicago, Illinois 60637**

If the tear-out card is missing or if you wish to place your order now, call (312) 363-5225 collect.

through the cusp pattern. Make the first line pass through both edges of the cusp but well away from the cusp point. The line cuts through two fold catastrophes, one catastrophe for each edge of the cusp pattern. Such a slice through the caustic pattern is said to be generic in the sense that a slight change in the line produces no major change in the type or number of catastrophes the path samples. Suppose I move the line slightly toward or away from the cusp point. It still passes over two edges and so still samples two fold catastrophes.

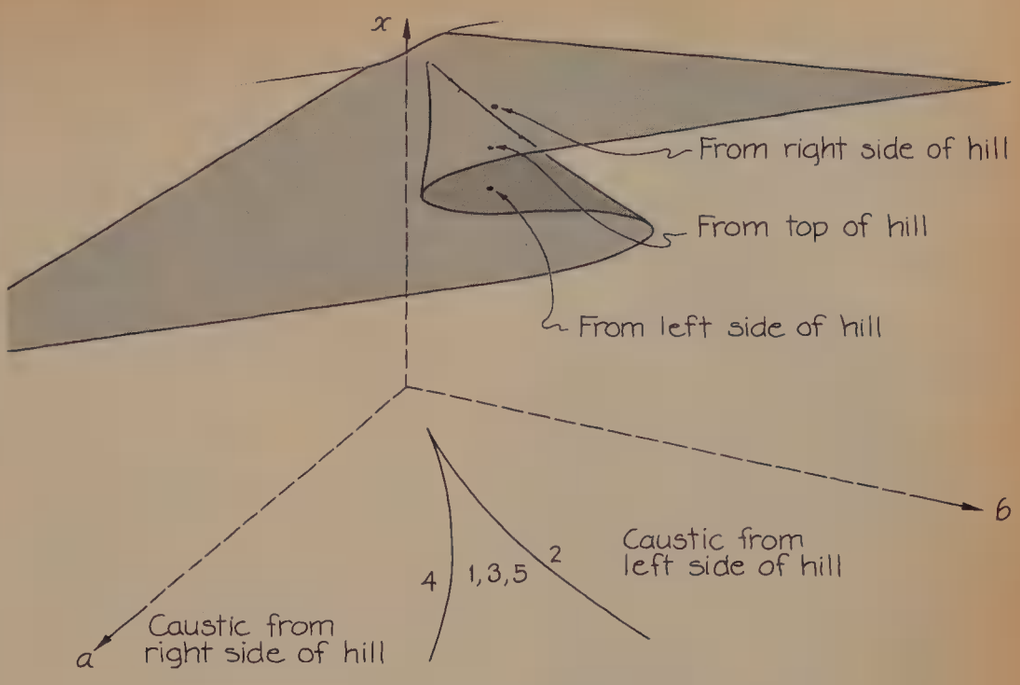
Next consider a line that passes directly through the cusp point. The line samples only one fold catastrophe, the one responsible for the point on the cusp itself. This slice through the catastrophe pattern is said to be nongeneric in the sense that a slight change in the line can have a dramatic effect on what the path samples. If I move the line slightly toward the rear of the graph, it no longer samples any of the catastrophe pattern. If I move the line slightly toward the front of the graph, it samples two fold catastrophes instead of one. Moving from a nongeneric section of a catastrophe pattern to a generic section is said to be unfolding the catastrophe.

The nature of the elementary catastrophes depends on the number of control and state variables involved in the production of a caustic pattern on the screen. To produce a fold catastrophe only one state variable (x) and one control variable (a) are required, but an additional control variable (b) is needed to produce a cusp catastrophe. If light passes through a refracting layer, many more control variables besides the two associated with a position on the screen are possible. For example, if the refracting material is a drop of water, a third control variable might be the pressure inside the drop, because the pressure can alter the shape of the drop.

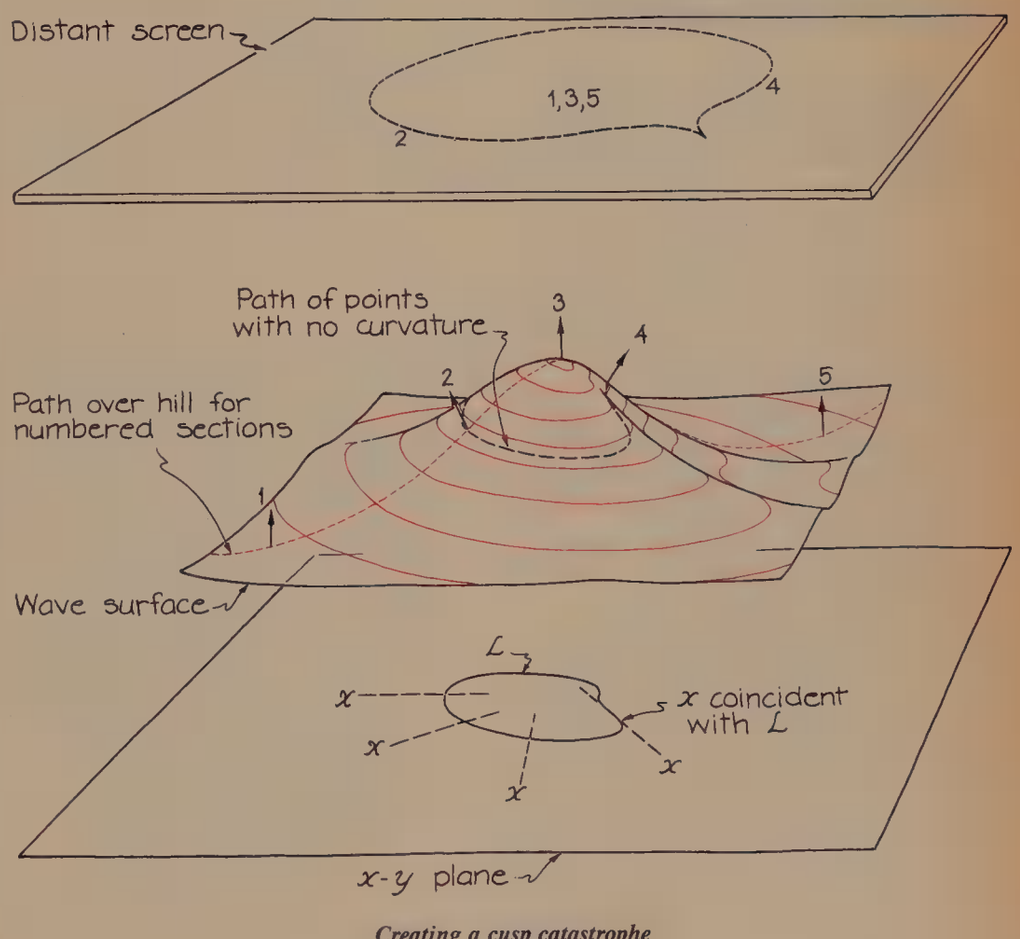
If a caustic pattern has one state variable (still the line x) and three control variables, the catastrophe is said to be a swallowtail one. The algebraic relation between the variables is obtained by differentiating the generating function with respect to the state variable and setting the result to zero. This relation cannot be graphed because it is four-dimensional.

The projection of the folded sheet onto a three-dimensional space of the control variables can be depicted. The procedure resembles what was done with the cusp catastrophe: the folded sheet was projected onto the bottom of the graph (the plane of the control variables) so that the cusp could be analyzed. The projection of the swallowtail folded sheet onto the control space is shown in the bottom illustration on the next page.

What is seen on the screen with such a catastrophe? It is certainly not the full projection (which is three-dimensional). Rather one sees only a cross-sectional slice through the projection. I have labeled the axes through the projection in order to facilitate making such slices. Two of the dimensions are the control variables a and b associated with places on the screen. The third variable (c) is some other variable one could change in the refraction demonstration.



A graph of a cusp catastrophe



Creating a cusp catastrophe

To visualize the patterns that can appear on the screen imagine a slice through the projection at one value for the third control variable. For example, a slice made near the left side of the illustration resembles a bird, which is the source of the name given to the catastrophe. If the third control variable has the proper value, this pattern can be seen on the screen.

Next imagine cutting slices through



Tina has never had a Teddy Bear.

A mother's love. A doll to cuddle. Tina knows nothing of these things. But she does know fear, rejection, and hunger.

For just $18 a month, you can help save a child like Tina.

Through our sponsorship program you can help provide a child with a better diet, clothes, medical attention, school. And even a toy or two.

But don't wait. There are so many. And somewhere, right now, a child is dying from starvation and neglect.

Write to: Mrs. Jeanne Clarke Wood Children, Incorporated, P. O. Box 5381, Dept. SA9T3, Richmond, Va. 23220 USA

☐ I wish to sponsor a boy ☐, girl ☐, in ☐ Asia, ☐ Latin America, ☐ Middle East, ☐ Africa, ☐ USA, ☐ Greatest Need.
☐ I will give $18 a month ($216 a year). Enclosed is my gift for a full year ☐, the first month ☐. Please send me the child's name, story, address and picture.
☐ I can't sponsor, but will help $_____
☐ Please send me further information.
☐ If for a group, please specify.

_____
Church, Class, Club, School, Business, etc.

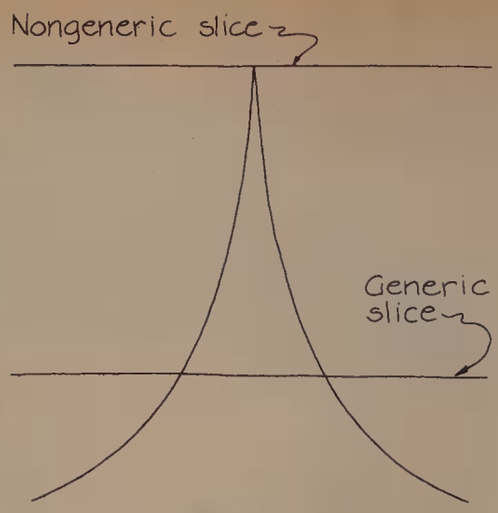
_____
NAME

_____
ADDRESS

_____
CITY STATE ZIP

U.S. gifts are fully tax deductible.
Annual financial statements are available on request.

CHILDREN, INC.



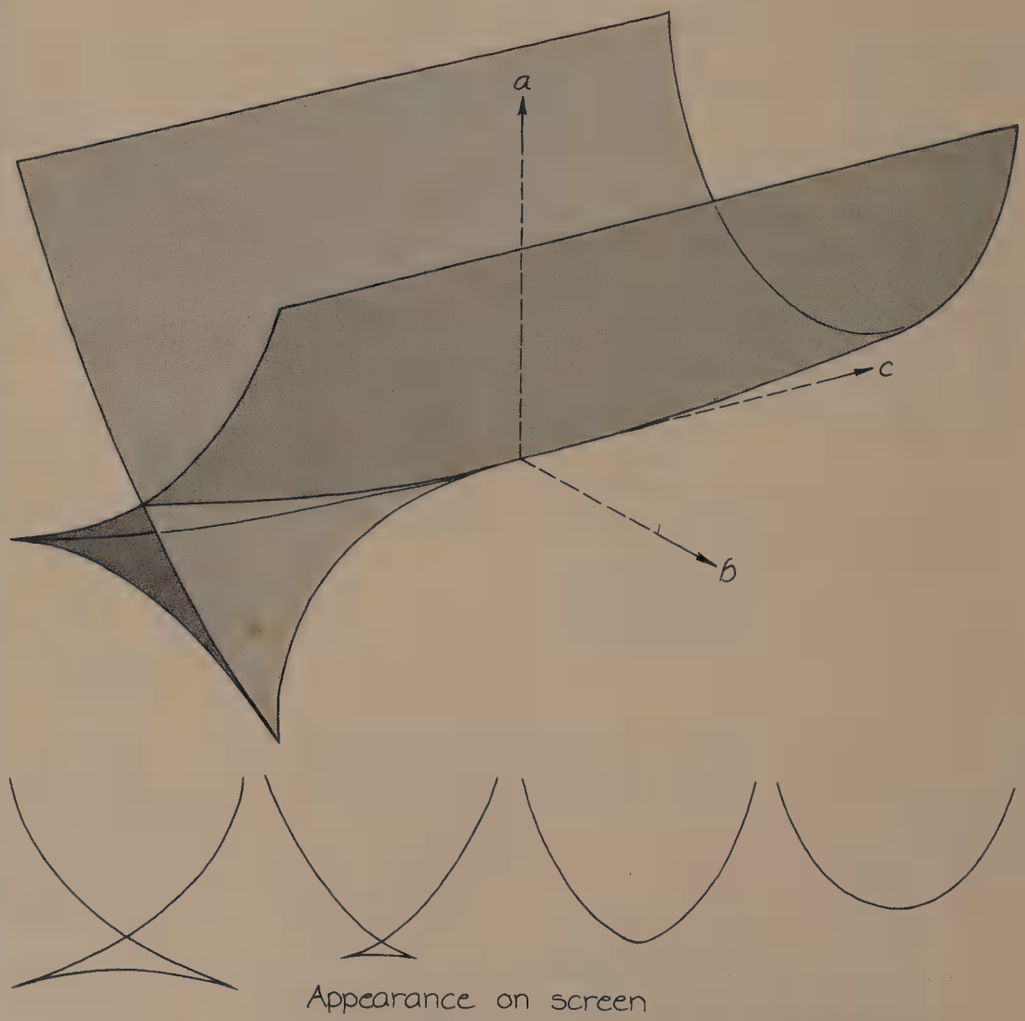
Slices through a cusp pattern

the projection for other values of the third control variable. The slices are made progressively toward the rear of the pattern. Eventually a slice has only a simple curve. Slices farther back are similar. Hence a slice from the front of the pattern yields a swallowtail on the screen and a slice from the rear produces a simple curve.

The slice dividing these two possibilities is a nongeneric section. The other slices through the projection are generic. If the third control variable can be controlled, one can vary the caustic pattern from the swallowtail through the smooth curve. Variations of this kind are described as folding or unfolding the catastrophe. For example, changing the slice from the nongeneric section to either the front or the back of the pattern is said to be unfolding the catastrophe.

When there are four control factors and one state factor, the catastrophe is of the butterfly type. The algebraic relation is obtained with the usual differentiating of the generating function. The result is impossible to graph because it represents a five-dimensional folded sheet. Even its projection onto the four-dimensional control space is impossible to draw. At best one can draw three-dimensional sections of the projection. I have not attempted the task but am content to provide the possible patterns that appear on the screen, which is only a two-dimensional slice out of those more complicated configurations. A few of the patterns of this type are shown in the top illustration on the opposite page. The catastrophe derives its name from the butterflylike shape of some of the patterns.

I have been considering the higher dimensions for the control variables while holding the control variable to x. A refracting material has another state variable, y. When there are three control variables and two state variables, the ca-



The swallowtail catastrophe

tastrophe is called a hyperbolic umbilic. The algebraic relation between these variables is obtained by differentiating the generating function with respect to x and then to y, with each result set to zero to produce two equations. Although the folded sheet representing these equations cannot be graphed, the projection of it onto the three-dimensional control space can be drawn as in the middle illustration at the right.

What appears on the screen is a slice through this pattern. The position of the slice is set by the value of the third control variable. The nongeneric slice is through the middle of the pattern. On the screen the caustic pattern is a bright corner with a finite angle (in contrast to the angle of zero degrees of a cusp corner). If one could actually control the third variable, this catastrophe could be unfolded by moving the slice. The pattern appearing on the screen would then change accordingly.

If the slice is moved toward the front or the back of the hyperbolic umbilic, the corner unfolds into a smooth curve lying around a cusp. Suppose such a pattern falls on the screen when a plastic layer is inserted into a laser beam. The pattern is said to be an unfolded hyperbolic umbilic.

A combination of three control variables and two state variables can also yield an elliptic umbilic. A different generating function is assigned and different algebraic equations are derived for the relation between the variables. The nongeneric slice provides a point caustic for the screen. Generic slices have triangles with curved sides. If the elliptic-umbilic catastrophe is unfolded, the caustic point turns into triangles.

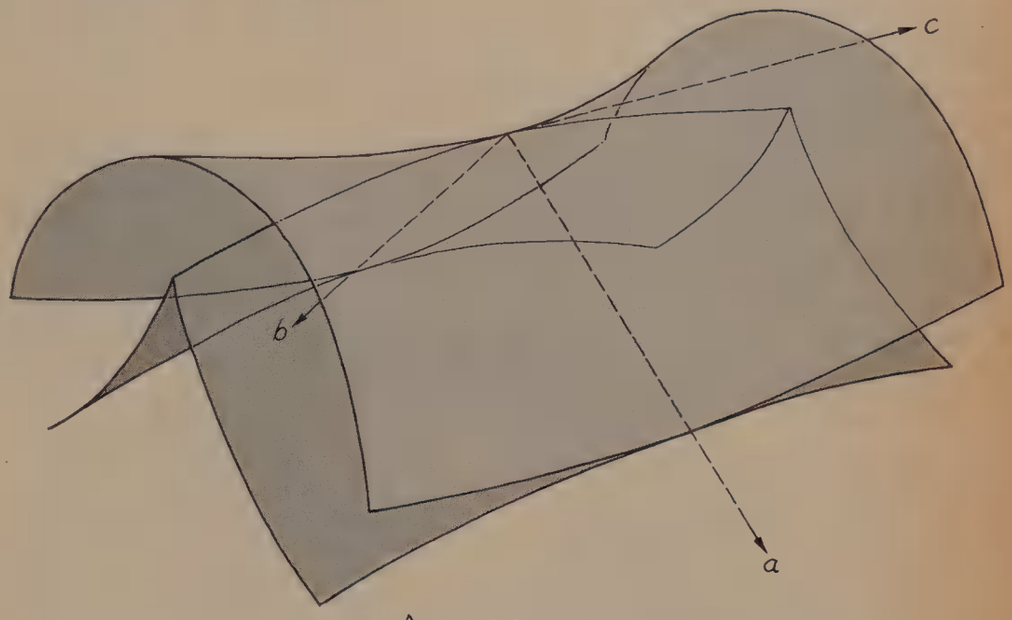
A further increase in the number of control variables for either one state variable or two is certainly possible, but the resulting higher-order catastrophes become difficult to understand. They are also unlikely. The chances are that only two or three control variables are important in any experiment with a refracting surface in a beam of light. Thus one should expect to find only the patterns from the first six elementary catastrophes on the screen.

You can do experiments with rippled plastic and other transparent materials for these catastrophe patterns. If you come on a pattern from a nongeneric slice through a catastrophe, you might try unfolding the catastrophe. I have had some success by carefully turning a plastic layer in a laser beam. Although I hold the same section of the plastic surface in the beam, the rotation of the plastic on an axis perpendicular to the beam forces the light to pass through the plastic at a new angle. The control variable I am changing is that angle. The result is that the higher-order catastrophe folds and unfolds.



Some of the patterns

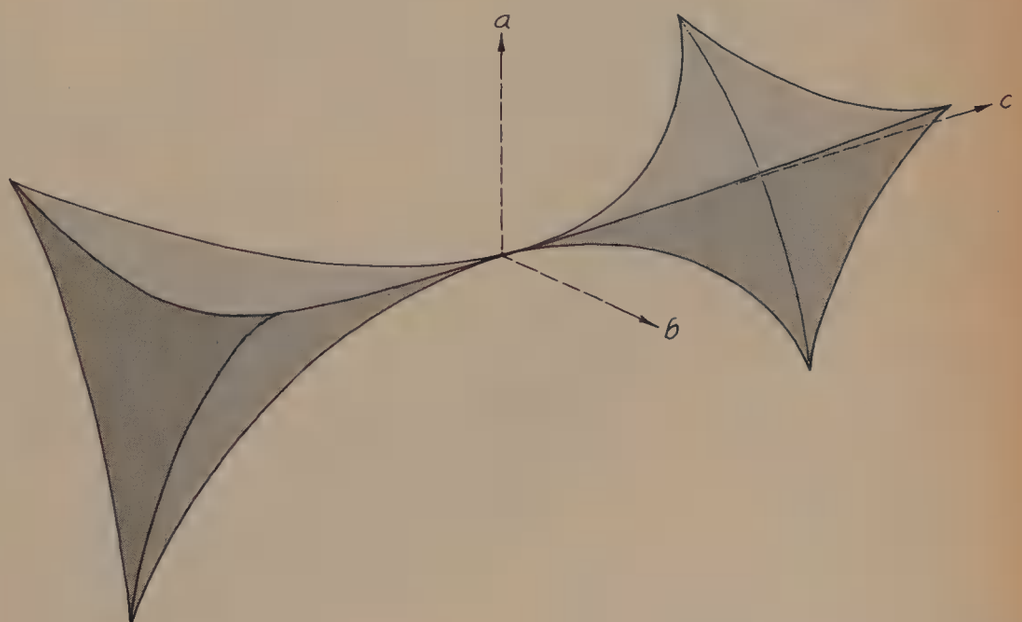
Butterfly catastrophes



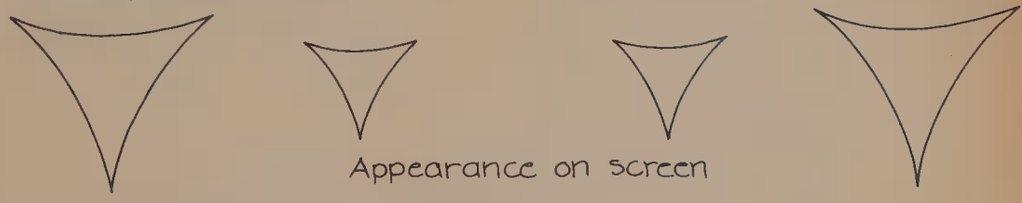
Appearance on screen



The hyperbolic-umbilic catastrophe



Appearance on screen



An elliptic-umbilic catastrophe

BIBLIOGRAPHY

Readers interested in further explanation of the subjects covered by the articles in this issue may find the following lists of publications helpful.

MATHEMATICAL GAMES

A QUICK TRIP THROUGH KNOT THEORY. R. H. Fox in *Topology of 3-Manifolds and Related Topics: Proceedings of the Topology of 3-Manifolds Institute, University of Georgia, 1961,* edited by M. K. Fort, Jr. Prentice-Hall, Inc., 1963.

AN ENUMERATION OF KNOTS AND LINKS, AND SOME OF THEIR ALGEBRAIC PROPERTIES. John Horton Conway in *Computational Problems in Abstract Algebra,* edited by John Leech. Pergamon Press, 1970.

KNOTS AND LINKS. Dale Rolfsen. Publish or Perish, Inc., 1976.

THE DYNAMIC EARTH

A REVOLUTION IN THE EARTH SCIENCES: FROM CONTINENTAL DRIFT TO PLATE TECTONICS. A. Hallam. Oxford University Press, 1973.

THE STEADY STATE OF THE EARTH'S CRUST, ATMOSPHERE AND OCEANS. Raymond Siever in *Scientific American,* Vol. 230, No. 6, pages 72–79; June, 1974.

ICE AGES: SOLVING THE MYSTERY. John Imbrie and Katherine Palmer Imbrie. Enslow Publishers, 1979.

LANGUAGE OF THE EARTH. Edited by Frank H. T. Rhodes and Richard O. Siever. Pergamon Press, 1981.

EARTH. Frank Press and Raymond Siever. W. H. Freeman and Company, 1982.

THE EARTH'S CORE

THE CURRENT STATUS OF SPECULATIONS ON THE COMPOSITION OF THE CORE OF THE EARTH. R. Brett in *Reviews of Geophysics and Space Physics,* Vol. 14, pages 375–383; 1976.

ORIGIN OF THE EARTH AND MOON. A. E. Ringwood. Springer-Verlag, 1978.

DRIVING MECHANISMS FOR THE EARTH'S DYNAMO. D. Gubbins and T. G. Masters in *Advances in Geophysics,* Vol. 21, pages 1–50; 1979.

BASALTIC VOLCANISM ON THE TERRESTRIAL PLANETS. Basaltic Volcanism Study Project. Pergamon Press, 1981.

MODELS OF THE EARTH'S CORE. D. J. Stevenson in *Science,* Vol. 214, No. 4521, pages 611–619; November 6, 1981.

THE INTERIOR OF THE EARTH: ITS STRUCTURE, CONSTITUTION AND EVOLUTION. M. H. P. Bott. Elsevier Science Publishing Co., Inc., 1982.

RECENT DEVELOPMENTS IN THE DYNAMIC THEORY OF PLANETARY MAGNETISM. F. H. Busse in *Annual Review of Earth and Planetary Sciences,* Vol. 11, pages 241–268; 1983.

THE EARTH'S MANTLE

CONVECTION IN THE EARTH'S MANTLE: TOWARDS A NUMERICAL SIMULATION. D. P. McKenzie, J. M. Roberts and N. O. Weiss in *Journal of Fluid Mechanics,* Vol. 62, Part 3, pages 465–538; February 11, 1974.

THE CHEMICAL ORIGIN OF THE EARTH'S MANTLE. R. K. O'Nions, P. J. Hamilton and Norman M. Evenson in *Scientific American,* Vol. 242, No. 5, pages 120–133; May, 1980.

PLANFORM OF MANTLE CONVECTION BENEATH THE PACIFIC OCEAN. Dan P. McKenzie, Anthony Watts, Barry Parsons and Micheline Roufosse in *Nature,* Vol. 288, No. 5790, pages 442–446; December 4, 1980.

THE OCEANIC CRUST

THE FACE OF THE DEEP. Bruce C. Heezen and Charles D. Hollister. Oxford University Press, 1971.

PLATE TECTONICS. Xavier Le Pichon and Jean Francheteau. Elsevier-Scientific Publishing Company, 1973.

PHOTOGRAPHIC ATLAS OF THE MID-ATLANTIC RIDGE RIFT VALLEY. Robert D. Ballard and James G. Moore. Springer-Verlag, 1977.

BIRTH OF AN OCEAN: THE CREST OF THE EAST PACIFIC RISE. Jean Francheteau, Thierry Juteau, David Needham and Claude Rangin. Centre National pour l'Exploitation des Océans, 1980.

THE CONTINENTAL CRUST

THE EVOLVING CONTINENTS. Brian F. Windley. John Wiley & Sons, Inc., 1977.

EVOLUTION OF THE EARTH'S CRUST. Edited by D. H. Tarling. Academic Press, 1978.

THE CONTINENTAL CRUST AND ITS MINERAL DEPOSITS. David W. Strangway. Special Paper No. 20, Geological Society of Canada, 1980.

STUDIES IN GEOPHYSICS: CONTINENTAL TECTONICS. Edited by B. C. Burchfiel, J. E. Oliver and L. T. Silver. National Academy of Sciences, 1980.

THE GROWTH OF WESTERN NORTH AMERICA. David L. Jones, Allan Cox, Peter Coney and Myrl Beck in *Scientific American,* Vol. 247, No. 5, pages 70–84; November, 1982.

THE OCEAN

TROPICAL RAINFOREST HISTORY AND THE EQUATORIAL PACIFIC CARBONATE DISSOLUTION CYCLES. N. J. Shackleton in *The Fate of Fossil Fuel CO_2 in the Oceans,* edited by Neil R. Andersen and Alexander Malahoff. Plenum Press, 1977.

DEEP-SEA CARBONATE AND THE DEGLACIATION PRESERVATION SPIKE IN PTEROPODS AND FORAMINIFERA. W. H. Berger in *Nature,* Vol. 269, No. 5626, pages 301–303; September 22, 1977.

GAIA: A NEW LOOK AT LIFE ON EARTH. J. E. Lovelock. Oxford University Press, 1979.

EXTRATERRESTRIAL CAUSE FOR THE CRETACEOUS-TERTIARY EXTINCTION. Luis W. Alvarez, Walter Alvarez, Frank Asaro and Helen V. Michel in *Science,* Vol. 208, No. 4448, pages 1095–1108; June 6, 1980.

TRACERS IN THE SEA. Wallace S. Broecker and T.-H. Peng. Lamont-Doherty Geological Observatory, Columbia University. Eldizia Press, Palisades, New York, 1982.

ICE CORE SAMPLE MEASUREMENTS GIVE ATMOSPHERIC CO_2 CONTENT DURING THE PAST 40,000 YR. A. Neftel, H. Oeschger, J. Schwander, B. Stauffer and R. Zumbrunn in *Nature,* Vol. 295, No. 5846, pages 220–223; January 21, 1982.

THE ATMOSPHERE

VARIATIONS IN THE EARTH'S ORBIT: PACEMAKER OF THE ICE AGES. J. D. Hays, John Imbrie and N. J. Shackleton in *Science,* Vol. 194, No. 4270, pages 1121–1132; December 10, 1976.

CLIMATIC CHANGE ON THE TERRESTRIAL PLANETS. James B. Pollack in *Icarus,* Vol. 37, No. 3, pages 479–553; March, 1979.

CLIMATE IN EARTH HISTORY. Wolfgang H. Berger and J. C. Crowell. National Academy Press, 1982.

THE BIOSPHERE

THE BIOSPHERE. G. Evelyn Hutchinson in *Scientific American,* Vol. 223, No. 3, pages 44–53; September, 1970.

COSMOS, EARTH, AND MAN: A SHORT HISTORY OF THE UNIVERSE. Preston Cloud. Yale University Press, 1978.

SYMBIOSIS IN CELL EVOLUTION: LIFE AND ITS ENVIRONMENT ON THE EARLY EARTH. Lynn Margulis. W. H. Freeman and Company, 1981.

THE EDIACARIAN PERIOD AND SYSTEM: METAZOA INHERIT THE EARTH. Preston Cloud and Martin F. Glaessner in *Science,* Vol. 217, No. 4562, pages 783–792; August 27, 1982.

THE AMATEUR SCIENTIST

WAVES AND THOM'S THEOREM. M. V. Berry in *Advances in Physics,* Vol. 25, No. 1, pages 1–26; 1976.

Marygrove College Library
Detroit, Michigan 48221
PLEASE DO NOT REMOVE

SCIENTIFIC AMERICAN



FUSION ENGINEERING

0 74820 08715 10

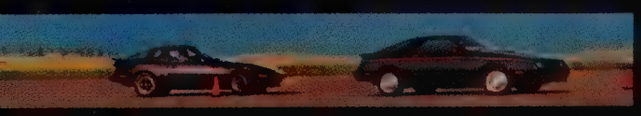
$2.50

October 1983

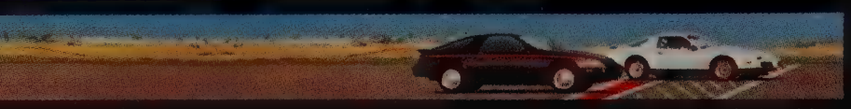
CHRYSLER CREATES LASER XE. The sports c

We gave Laser XE world-class performance. In the slalom, Laser beats all entries—from Trans Am to Mustang GT.

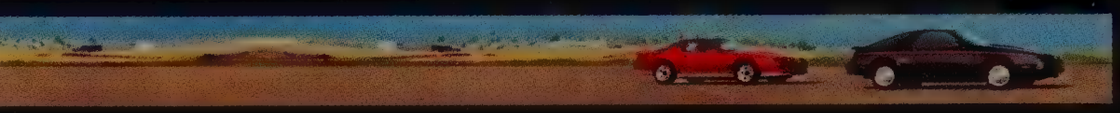
We built Laser XE to outperform the competition: Camaro Z28, Trans Am, Mustang GT, Toyota Supra, Mazda RX-7. Laser does it when you equip it with turbo, performance handling package and nitrogen-charged shocks.* Laser does it with front-wheel drive, new dual-path suspension system and quick-ratio power steering. In the slalom Laser finishes No. 1—even ahead of the legendary Porsche 944.



Laser beats Porsche in the slalom.



Laser outperforms Trans Am in braking.



Laser is faster than Camaro Z28 from 0-50 mph.



We turbocharged it. From 0-50, Laser XE leaves Camaro Z28 with its shadow.

Z28's powerhouse but Laser XE's the sophisticated new wave. Its multi-point injection system "spritzes" fuel in at four points. Its water-cooled bearing reduces critical turbo temperature by 500°F. Its turbo engine boosts h.p. 45% and moves Laser like light.

With 5-speed your time to 50 mph is 5.8 seconds. Z28, Trans Am, Supra and RX-7 are your remote-controlled side-view mirrors.

We gave it high-performance braki Laser XE stops where Trans Am doe

We think total performance calls formance braking. So we gave Lase semi-metallic brake pads, power bra around and optional wide 15" all wheels with Goodyear Eagle GT radial tires. Result: Laser stops quicker than Z28, Trans Am,



Established 1845 **SCIENTIFIC AMERICAN** October 1983 Volume 249 Number 4

ARTICLES

54 **LABOR-INTENSIVE AGRICULTURE, by Philip L. Martin**
The tardy mechanization of the U.S. fruit and vegetable industry invites competition from abroad.

60 **THE ENGINEERING OF MAGNETIC FUSION REACTORS, by Robert W. Conn**
Work on fusion power is going beyond scientific studies of feasibility to engineering studies.

72 **THE LARGE-SCALE STRUCTURE OF THE UNIVERSE, by Joseph Silk, Alexander S. Szalay and Yakov B. Zel'dovich** It may reflect perturbations in density early in the big bang.

90 **THE PROCESSING OF RNA, by James E. Darnell, Jr.**
In cells with a nucleus a lot happens to messenger RNA after it has been transcribed from DNA.

102 **THE DEAD SEA, by Ilana Steinhorn and Joel R. Gat**
Five years ago the world's saltiest lake "turned over," dissipating an isolated body of fossil water.

120 **SIX MILLENNIUMS OF BUFFALO KILLS, by B. O. K. Reeves**
Indians stampeded herds of buffalo over a cliff at a site in western Canada starting in 3700 B.C.

136 **THE EXTINCTION OF THE AMMONITES, by Peter Ward**
These animals resembling the nautilus suddenly died out some 65 million years ago. Why?

149 **BILATERAL NEGOTIATIONS AND THE ARMS RACE, by Herbert F. York**
A historical review of the talks between the U.S. and the U.S.S.R. explores their basic blocks.

DEPARTMENTS

6 LETTERS
11 50 AND 100 YEARS AGO
12 THE AUTHORS
22 COMPUTER RECREATIONS
51 BOOKS
82 SCIENCE AND THE CITIZEN
162 THE AMATEUR SCIENTIST
170 BIBLIOGRAPHY

BOARD OF EDITORS Gerard Piel (Publisher), Dennis Flanagan (Editor), Brian P. Hayes (Associate Editor), Philip Morrison (Book Editor), John M. Benditt, Peter G. Brown, Michael Feirtag, Diana Lutz, Jonathan B. Piel, John Purcell, James T. Rogers, Armand Schwab, Jr., Joseph Wisnovsky

ART DEPARTMENT Samuel L. Howard (Art Director), Steven R. Black (Assistant Art Director), Ilil Arbel, Edward Bell

PRODUCTION DEPARTMENT Richard Sasso (Production Manager), Carol Eisler and Leo J. Petruzzi (Assistants to the Production Manager), Carol Hansen (Electronic Composition Manager), Carol Albert, Karen Friedman, Lori Mogol, Karen O'Connor, Julio E. Xavier

COPY DEPARTMENT Sally Porter Jenks (Copy Chief), Debra Q. Bennett, Mary Knight, Dorothy R. Patterson

GENERAL MANAGER George S. Conn

ADVERTISING DIRECTOR C. John Kirby

CIRCULATION MANAGER William H. Yokel

SECRETARY Arlene Wright

SCIENTIFIC AMERICAN (ISSN 0036-8733), PUBLISHED MONTHLY BY SCIENTIFIC AMERICAN, INC., 415 MADISON AVENUE, NEW YORK, N.Y. 10017. COPYRIGHT © 1983 BY SCIENTIFIC AMERICAN, INC. ALL RIGHTS RESERVED. PRINTED IN THE U.S.A. NO PART OF THIS ISSUE MAY BE REPRODUCED BY ANY MECHANICAL, PHOTOGRAPHIC OR ELECTRONIC PROCESS, OR IN THE FORM OF A PHONOGRAPHIC RECORDING, NOR MAY IT BE STORED IN A RETRIEVAL SYSTEM, TRANSMITTED OR OTHERWISE COPIED FOR PUBLIC OR PRIVATE USE WITHOUT WRITTEN PERMISSION OF THE PUBLISHER. SECOND-CLASS POSTAGE PAID AT NEW YORK, N.Y., AND AT ADDITIONAL MAILING OFFICES. AUTHORIZED AS SECOND-CLASS MAIL BY THE POST OFFICE DEPARTMENT, OTTAWA, CANADA, AND FOR PAYMENT OF POSTAGE IN CASH. SUBSCRIPTION RATE: $24 PER YEAR IN THE U.S. AND ITS POSSESSIONS; $33 PER YEAR IN ALL OTHER COUNTRIES. POSTMASTER: SEND ADDRESS CHANGES TO SCIENTIFIC AMERICAN, 415 MADISON AVENUE, NEW YORK, N.Y. 10017.

SCIENTIFIC AMERICAN

CORRESPONDENCE

Offprints of more than 1,000 selected articles from earlier issues of this magazine, listed in an annual catalogue, are available at 75 cents each. Correspondence, orders and requests for the catalogue should be addressed to W. H. Freeman and Company, 4419 West 1980 South, Salt Lake City, UT 84121. Offprints adopted for classroom use may be ordered direct or through a college bookstore. Sets of 10 or more Offprints are collated by the publisher and are delivered as sets to bookstores.

Photocopying rights are hereby granted by Scientific American, Inc., to libraries and others registered with the Copyright Clearance Center (CCC) to photocopy articles in this issue of SCIENTIFIC AMERICAN for the flat fee of $2 per copy of each article or any part thereof. Such clearance does not extend to the photocopying of articles for promotion or other commercial purposes. Correspondence and payment should be addressed to Copyright Clearance Center, Inc., 21 Congress Street, Salem, MA 01970. Specify CCC Reference Number ISSN 0036–8733/83. $2.00 + 0.00.

Editorial correspondence should be addressed to The Editors, SCIENTIFIC AMERICAN, 415 Madison Avenue, New York, NY 10017. Manuscripts are submitted at the authors' risk and will not be returned unless they are accompanied by postage.

Advertising correspondence should be addressed to C. John Kirby, Advertising Director, SCIENTIFIC AMERICAN, 415 Madison Avenue, New York, NY 10017.

Subscription correspondence should be addressed to Subscription Manager, SCIENTIFIC AMERICAN, P.O. Box 5969, New York, NY 10017. The date of the last issue on your subscription is shown in the upper right-hand corner of each month's mailing label. For change of address notify us at least four weeks in advance. Please send your old address (if convenient, on a mailing label of a recent issue) as well as the new one.

Name

New Address

Street

City

State and ZIP

Old Address

Street

City

State and ZIP



THE COVER

The painting on the cover shows a pair of superconducting magnets in what is called a yin-yang configuration. They form a magnetic mirror, that is, they trap charged particles such as atomic nuclei by reflecting them between peaks of magnetic field strength; thus they symbolize the attempts to confine a plasma (a gas of charged particles) magnetically and then heat it so that thermonuclear fusion is possible (see "The Engineering of Magnetic Fusion Reactors," by Robert W. Conn, page 60). The configuration, which is about eight meters long, was designed and built by the General Dynamics Corporation and the Lawrence Livermore National Laboratory. It was tested at Livermore, where it produced a maximum field strength of some 80,000 gauss. It is currently being incorporated into a larger device at Livermore, a tandem mirror, which requires a yin-yang configuration at each end of a linear sequence of magnets.

THE ILLUSTRATIONS

Cover painting by Ted Lodigensky

Page	Source
27–36	Ilil Arbel
55	Barrie Rokeach
56–59	Ilil Arbel
60	Ian Worpole
62–69	Gabor Kiss
70	F. W. Wiffen, J. O. Steigler and E. E. Bloom, Oak Ridge National Laboratory
71	Gabor Kiss
73	John E. Moody, Edwin L. Turner and J. Richard Gott III, Princeton University
74	S. Djorgovski, University of California at Berkeley
75–77	Walken Graphics
78	Carlos Frenk and Simon White, University of California at Berkeley; Marc Davis, Center for Astrophysics of the Harvard College Observatory and the Smithsonian Astrophysical Observatory
79	Walken Graphics
80	George Efstathiou, University of Cambridge
90–91	Ilil Arbel
92–99	Jerome Kuhl
103	EROS Data Center
104–109	J. Dyke Fledderus
121	B. O. K. Reeves, University of Calgary
122–124	Andrew Tomko
124D–124H	B. O. K. Reeves, University of Calgary
128–129	Patricia J. Wynne
130	Ilil Arbel
135	Patricia J. Wynne
137	William K. Sacco, courtesy of Peabody Muesum of Natural History, Yale University
138–146	Tom Prentiss
150–153	Ilil Arbel
162–168	Michael Goodman



Sleep Sheep

We may have found a way to improve your sleep forever.

New Zealanders discovered that sleeping under sheep's wool induced sleep.

The story we are about to tell you may seem rather incredible. And indeed it is. But if you'll have an open mind, what you will learn may indeed change your life.

There is a new product manufactured in New Zealand that is selling very well. It's called the Woolrest—a bed pad made of thick wool. You simply place it over your mattress and then cover it with your bottom sheet as you normally do when you make your bed.

SLEEP INDUCING

Sheep's wool has always had an outstanding reputation for keeping you warm in winter and cool in summer. That is why wool seat covers are so popular. But wool has another property known by New Zealanders for many years. Namely—wool induces sleep.

In New Zealand, for example, a way to cure insomnia was to cover yourself with a sheepskin rug. For some very unscientific reason, the hypnotic quality of the wool encouraged sleep. Counting sheep was another technique that was often recommended. But it was the development of the Woolrest bed pad that suddenly turned an old wive's tale into fact.

DIFFERENT EXPERIENCE

Sleeping on a Woolrest is a different sleeping experience. Whether you sleep on a hard or soft mattress, the Woolrest apparently radiates a feeling of comfort and relaxation from the wool fibers which mold, massage, and conform to your entire body.

Scientific tests conducted by Dr. Peter Dickson of Ohio State University proved that the Woolrest pad indeed helped induce sleep. Testimonials from people who owned them clearly demonstrated that Woolrest not only induced sleep but also provided great relief for backaches, arthritis, and rheumatism. Its natural fibers tend to alleviate the pain and pressure caused by these illnesses and thus make sleep come easier and deeper.

Frank Thornton, a Seattle Washington dental technician, had trouble sleeping for six years. He purchased a Woolrest pad and has been able to sleep through the entire night practically every night. According to Thornton, "I felt a sensation of buoyancy from the Woolrest. I have read in literature that a person sleeps more restfully with natural fibers surrounding their body." Thornton has already purchased seven Woolrest pads for his entire family.

There are hundreds of other testimonials we could mention from the thousands who have purchased them, but one thing is clear. Using a Woolrest pad you know that the rest of your life will be spent sleeping comfortably, with a minimum amount of sleeplessness. Certainly, there will be nights when it will be difficult to go to sleep. But with the Woolrest those nights will be fewer, shorter and without the tradeoff of taking sleep-inducing drugs.

WARM IN WINTER

In the winter the Woolrest holds your body heat and thus keeps you warmer. You'd expect that. But in summer it keeps you cooler through a process of moisture absorption by absorbing up to one-third of its weight in moisture to keep your body cool and dry.

Will the Woolrest work for you? We're willing to prove it with no risk on your part. Order one for a 45-day sleep test. When you receive it, closely examine the surface. Feel the thick pure woven pile of natural New Zealand wool. Place it on your bed and then cover the pad with your regular bottom bed sheet—either fitted or plain.

JUST A FEW DAYS

That night go to sleep on it. But don't judge the effectiveness until you've slept on it for at least a full month. You should actually notice the difference in just a few days, but after a month you'll turn from being skeptical like we were, to a real enthusiastic believer. If not, we'll understand. After all it may not work for everybody. Just return it within the 45-day sleep trial and you'll receive a prompt and courteous refund including the $4 delivery charge. The Woolrest washes easily in your washing machine and has a limited five-year warranty. Complete instructions come with each pad. Sleep on a Woolrest and experience what we mean at no risk or obligation. Order one, today.

To order, credit card holders call toll free and ask for product by number below or send check plus $4 delivery for each pad ordered.

Twin Size (0025F)	**$129**
Full Size (0026F)	**179**
Queen Size (0027F)	**209**
King Size (0028F)	**269**

The New Zealand wool used exclusively in the Woolrest could be a major reason for its effectiveness. Here's why.

Most sheep are raised in countries where temperature extremes cause their wool fibers to become brittle, rigid and often cracked at the ends. In Australia sheep are often raised in the dusty outback where temperature and rainfall can vary greatly.

In New Zealand however, the sheep enjoy a year-round temperate climate and graze in lush green pastures. Their wool fibers are thick and vibrant with good shape retention .

NEW ZEALAND WOOL

The Woolrest uses only the best New Zealand wool. The fibers are actually woven onto a pure wool blanket base and thoroughly washed and tufted.

We have mentioned the importance of New Zealand wool for good reason. Just as there are several grades of sheepskin seat covers, we suspect eventually there will be several grades of wool bed pads. None will ever compare to the wool used in the Woolrest.

JS&A®

One JS&A Plaza
Northbrook, Illinois 60062

CALL TOLL FREE 800 323-6400

IL residents add 6% sales tax. ©JS&A Group, Inc.,1983



LEARN TO SPEAK
SPANISH
WHILE DRIVING TO WORK
WITH THE AMR® PATENTED
BI-LINGUAL
STEREO-SELFTEST™ METHOD
YOU CAN NOT HELP BUT PICK UP
A SECOND LANGUAGE
WITHOUT OPENING A BOOK
SPANISH-FRENCH-GERMAN-ITALIAN
CHINESE-JAPANESE-ENGLISH
Call us at AMR and a staff member will gladly answer your questions
☎ 1-800-323-2500 ☎
IL Residents 312-858-1900
AMR® EDUCATIONAL SYSTEMS
800 Roosevelt Rd.
Glen Ellyn, Illinois
60137

THE
HEBREW
UNIVERSITY
OF
JERUSALEM

Jerusalem, Israel

The Hebrew University of Jerusalem offers a small number of Post-doctoral Fellowships for the 1984-85 academic year in the Humanities, Social Sciences, Natural Sciences, Agriculture and Medicine.

Candidates may apply at an early stage of their professional career (not later than 3 years after completion of their doctoral dissertation) to P.O. Box 1255, Jerusalem 91904, Israel.

Completed application forms must be returned by December 1, 1983.

LETTERS

Sirs:

In "The Slowing of Urbanization in the U.S." [SCIENTIFIC AMERICAN, July] Larry Long and Diana DeAre write: "The term Megalopolis was coined by Jean Gottman of the University of Oxford in a book of the same name published in 1961." I don't think so. Gottman may have been the first to apply it to the East Coast corridor between Boston and Richmond, but he was certainly not the first to employ it to categorize the emerging supercities (in population and area) of the late 20th century.

In *The Decline of the West,* Oswald Spengler (at least in the translation of Charles Francis Atkinson) uses the word (in both its nominal and adjectival forms) and discusses its significance at some length and depth (Vol. I, Chapter 1, and Vol. II, Chapter 4).

Of some interest is the fact that there was a city named Megalopolis in ancient Greece. It was in the Peloponnesus a few miles northwest of Sparta. It can be found on any good map of ancient Greece and was mentioned by many of the classical authors. Because Megalopolis was not remarkable for its size, it would be interesting to know how it got its name.

RORY FORAN

Glen Burnie, Md.

Sirs:

In "The Salmonid Fishes as a Natural Livestock" [SCIENTIFIC AMERICAN, July], Lauren R. Donaldson and Timothy Joyner have given a thorough account of advances in the nutrition and breeding of salmon. This report of their contribution to our need for food from the sea is indeed significant.

The authors graciously acknowledge our discovery of home-stream olfactory imprinting in salmon, but the article they cite ["The Homing Salmon," by Arthur D. Hasler and James A. Larsen; SCIENTIFIC AMERICAN, August, 1955] was written when that discovery was still a hypothesis. It was not until 1976 that our group conducted the rigorous studies using the synthetic odors morpholine and phenethyl alcohol as imprinters on salmon introduced into Lake Michigan that the hypothesis was proved. (See "Imprinting to Chemical Cues: The Basis for Home Stream Selection in Salmon," by Allan T. Scholz, Ross M. Horrall, Jon C. Cooper and Arthur D. Hasler in *Science,* Vol. 192, No. 4245, pages 1247–1249, June 18, 1976, and "Olfactory Imprinting and Homing in Salmon," by Arthur D. Hasler, Allan T. Scholz and Ross M. Horrall in *American Scientist,* Vol. 66, No. 3, pages 347–355, May–June, 1978.)

Donaldson and Joyner find it pointless to imprint salmon with synthetic chemicals. It is this technique, however, that has improved the harvest of salmon in Lake Michigan. On the Wisconsin side there are no streams in which salmon can reproduce, hence salmon must be artificially reared in local hatcheries. In some cases the young hatchery salmon can be stocked in smolting ponds next to a stream where natural imprinting can occur. It is often advantageous, however, to imprint the salmon artificially in a hatchery and then stock them directly into the lake. When these salmon mature, they have no home stream, only a "memory" of the "home" hatchery. They can therefore be decoyed to the best and most convenient harvest sites by scenting the inlets of selected streams with imprinting chemicals. An example is morpholine, which was added to the hatchery water during the salmon's smolting period.

ARTHUR D. HASLER

ROSS M. HORRALL

ROBERT A. RAGOTZKIE

University of Wisconsin
Madison

Sirs:

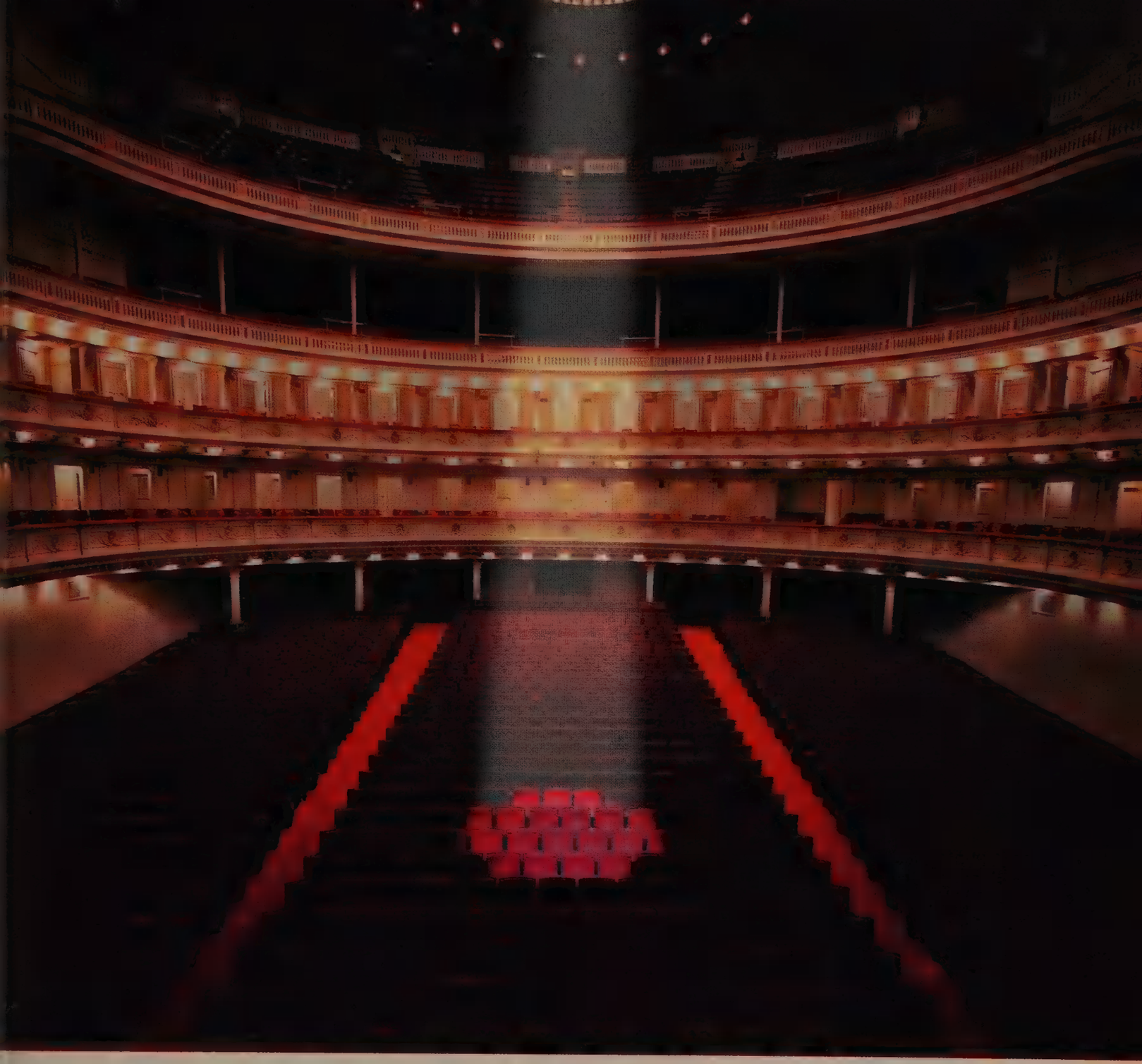
Albert J. Blodgett, Jr.'s, "Microelectronic Packaging" [SCIENTIFIC AMERICAN, July] was an excellent introduction to the subject. In focusing on IBM's water-cooled thermal-conduction module, however, he overlooked the more advanced air-cooled packaging used in Amdahl Corporation's 580 series of large computers.

Each Amdahl chip is bonded to a chip carrier whose upper surface has a metal stud with numerous circular cooling fins, which transfer the chip's heat to the airstream. Up to 121 logic and memory chips are mounted on a ceramic multichip carrier (MCC) similar to the substrate for IBM's TCM. The entire central processing unit of a fully configured 580 consists of a stack of nine MCC's connected by printed circuit boards. It occupies 5.6 cubic feet.

Compared with a water-cooled design, air-cooled packaging reduces system complexity, manufacturing costs, power consumption and floor space needed, while increasing reliability and ease of maintenance. The 580's advanced packaging is also one reason it has a better price/performance ratio than the 3081.

GARY HLADIK

Saratoga, Calif.



Sony creates seventh row, center. Forever.

INTRODUCING THE SONY COMPACT DISC PLAYER.

You are looking at an invention so extraordinary that High Fidelity magazine has hailed it as "the most fundamental change in audio technology in more than eighty years."

A piece of audio equipment that permits you to hear something you've never heard before: perfection.

Sony is pleased to raise the curtain on the world's first digital audio compact disc player. The CDP-101.

THE END OF HI-INFIDELITY.

The CDP-101 uses a laser beam to play compact discs that hold far more musical information than conventional records ever could. The result: fidelity and depth of sound that no conventional turntable could ever deliver.

This highest of fidelity remains faithful too. Because the digital discs are read by laser beam, there is absolutely no physical wear. Nor is there any distortion from such annoyances as dust, fingerprints or even scratches. The laser simply sees through them.

Equally ingenious, an infrared remote control lets you select tracks without budging from your armchair.

While an ever-expanding library of compact discs lets you listen to your favorite artists as though you, and your armchair, were centered in the spotlight above.

Maybe most important, the CDP-101 is fully compatible with the conventional equipment you may own.

We suggest you hear the Sony Compact Disc Player soon. For a sound you can't believe, from the audio innovator you assuredly can.

SONY®
THE ONE AND ONLY



© 1983 Sony Corp. of America. Sony is a registered trademark of the Sony Corp. 1 Sony Drive, Park Ridge, New Jersey 07656.

Killer pigmy

INTRODUCING A TOTAL COMPUTER SYSTEM NO LARGER THAN THIS MAGAZINE.

You're looking at a miracle. A creation of technology that you can actually use.

The amazing Epson Notebook Computer.®

Go ahead – put your hands on that full-sized keyboard. You now control a *complete* computer system. A system that includes a liquid crystal display screen, a micro cassette document and program storage unit, a 50-hour rechargeable internal power supply, and a dot matrix printer. All for *less* than $800.



Nothing else can equal it. Not the 30-pound "transportables" that are like lugging sewing machines around, and certainly not those so-called "portables" that try to skimp by with no printer, no provision for document storage, and a maddeningly short battery life.

The Epson Notebook Computer has everything you need to get down to work. Exquisitely crafted into a sleek and silent package that you can take anywhere and use anywhere.

It weighs less than four pounds. Has a surface area the size of a sheet of typing paper. And it contains a whopping 16K of random memory (optionally expandable to 32K, if you need it).

You can program the Notebook Computer in a built-in, extended version of Microsoft® BASIC. An internal word processing program called SkiWriter® will have you turning out letters, notes, memos, or full reports after just fifteen minutes of reading the simple instructions.

Available software ranges from personal productivity programs to education and entertainment. We even offer a telephone modem and software that allows you to instantly transmit information to a full-sized computer.

To put your hands on a *real* Epson Notebook Computer, call toll-free (800) 421-5426, or (213) 539-9140 in California for the name of your nearest Epson dealer.

But be warned. Few who touch the "killer pigmy" go away empty-handed.



EPSON

STATE-OF-THE-ART. . . SIMPLICITY.

actual size:



Microsoft is a trademark of Microsoft Corp. SkiWriter is a trademark of SkiSoft Inc.



TWO TOUGH OUTFITS!
FORD BRONCO II AND EDDIE BAUER.

SPECIAL EDDIE BAUER BRONCO II

Two names known for toughness get together –in Ford's special Eddie Bauer Bronco II.

Like all Bronco's, Bronco II is field-tested ...built Ford tough, with husky frame and Twin-Traction-Beam suspension. Eddie Bauer survival gear has been field-tested too...right to the top of Mount Everest!

The Eddie Bauer Bronco II is outfitted for all outdoors–with unique interior and exterior trim plus an optional flip-away spare tire carrier. Free Eddie Bauer top quality gear bag and map folder included, plus scheduled maintenance and a limited warranty for 24 months/24,000 miles, whichever comes first, excludes tires, fluids, accidents or abuse.

Trimmer Bronco II is fun to drive, easy to park and maneuver...with a V-6 engine standard. Seats 4 in comfort, with split fold-down rear seats for extra cargo.

Eddie Bauer Bronco II –it's a brand-new kick in survival gear!

Bronco II

Eddie Bauer

Ford

Get it together–Buckle up!

50 AND 100 YEARS AGO

SCIENTIFIC AMERICAN

OCTOBER, 1933: "The Tennessee Valley Authority Act of 1933 has granted President Roosevelt the power to appoint a body of three members to develop the natural resources of the Tennessee River basin, with special reference to the improvement of navigation, which would be linked to flood control; reforestation, which would embrace improvement of the unemployment situation; utilization of marginal lands, which have been overworked for agricultural purposes, and finally the development of great blocks of power for the further industrialization of the entire Tennessee River Valley area. Admitting that this would be a gigantic Federal experiment on a scale never heretofore undertaken, the President has suggested the possibility of carrying the idea further into other regions if the Tennessee River basin project indicated that it would succeed. The area under the jurisdiction of the Tennessee Valley Authority covers most of central and western Tennessee, extends widely over northern Alabama, includes portions of North Carolina, South Carolina and Georgia, and touches small portions of Virginia, West Virginia, Kentucky and Mississippi."

"There is a very satisfactory theory of what happens in a nova—except for the greatest question of all. What causes the tremendous initial outburst? A noteworthy suggestion has been made by E. A. Milne, namely that the main inner part of a star may become unstable as regards its internal equilibrium and suddenly collapse to a state of much greater density and smaller diameter. It is possible that in such an event practically the whole of the enormous amount of gravitational energy liberated by the contraction might be stored up in one form or another in a collapsed mass, leaving but a relatively small amount to be got rid of by radiation. The new and smaller core, however, would at first be exceedingly hot. The rapid progress of nuclear physics makes it reasonable to hope that, before many years have passed, we may know enough to say whether such suggestions are to be accepted."

"The Pennsylvania Railroad is engaged in what is one of the greatest, if not the greatest, single electrification projects ever undertaken in this country or any other. It consists in the electrification of 1,082 miles of railroad track, serving the great cities of the Eastern Seaboard from New York to Washington and extending westward to the Susquehanna River in the heart of Pennsylvania. It is expected that 816,000,000 kilowatt-hours per year will be required for the operation of trains when the project is completed. The project's magnitude and its possibilities are great enough to stir the imagination of the most prosaic."

"It is being asserted widely that science has at last abandoned its early materialistic taint, swinging strongly toward religion and mysticism. The furor started small in a corner of physical science. It was discovered a few years ago that things as small as electrons do not individually obey the law of cause and effect. This phenomenon is known as the principle of indeterminacy, or the principle of uncertainty, and its discovery came about when the German physicist Werner Heisenberg showed that it is impossible to ascertain both the position and speed of an electron; we can ascertain the one or the other singly, but not both. Evidently, then, the behavior of an electron is indeterminable. Right here a number of thinkers made false deductions and, as A. S. Eddington put it, 'science went off the gold standard.' What these thinkers failed to grasp was that mere indeterminability does not in itself establish indeterminacy. A thing may be indeterminable but not indeterminate. Nature knows what she is doing, and does it, even when we cannot find out. It did not take the mystics long to discover the principle of uncertainty. If we could no longer predict, at least in theory, the entire future of the universe, given the position and velocity of every particle in it, then perhaps there was freedom in it after all. The return of science to some sort of modern mysticism would be essentially a slip in man's hard-won progress away from one of his most ancient bad habits—that of ascribing to the supernatural whatever he did not yet understand."



OCTOBER, 1883: "The Straits of Sunda, separating the islands of Java and Sumatra, form one of the main gateways used by the vast trade that navigates the China Sea. Half-way through the strait, equidistant from the two shores, was a group of three islands, the largest of which was Krakatoa, four and a half miles long and three miles broad, its volcanic summit reaching to a height of 2,623 ft. above the sea level, about 10 times higher than the surrounding sea was deep. The tremendous volcanic eruption, with the accompanying earthquake and inundation of the coasts, on Sunday, the 26th of August, has wrought a fearful change. As far as our present information goes, it appears that the whole island of Krakatoa, consisting of about 8,000 million cubic yards of material, has fallen in and disappeared below the sea."

"Apart from oxygen and nitrogen there is but one other substance in dry air that we are at present warranted in regarding as being a necessary and constant component, namely carbonic acid or carbon dioxide (CO_2). Small as its proportion in the air is, its relation to animal and vegetable life on the earth has long been recognized. The gas is liberated in abundance by the action of heat on limestone and other carbonates, and also by the spontaneous decomposition of solutions of bicarbonate of lime. The abundant deposits of limestone in the crust of the earth form, therefore, an inexhaustible source of the gas under certain conditions, and their abundance points probably to a period in the earth's history when a much higher proportion of carbonic acid was present in the air. While all the evidence goes to show that carbonic acid is now an almost invariable constituent of the air, it is one that requires least change in the physical conditions under which the earth exists to effect a change in its proportion. Minute as the proportion is, the delicacy of its relation to animal and vegetable life on the earth makes the maintenance of the apparently unstable equilibrium a matter of serious concern to mankind."

"Professor Albert S. Bickmore describes his methods of imparting instruction in natural history by ocular demonstration, 'believing that the sense of sight is the royal avenue to the mind.' A large part of the objects that it was desired to display were either too small or too large to be taken to the lecture hall, and at the same time too important to be omitted. To overcome this difficulty the most complete stereopticon to be found was purchased; and as it was discovered that photographic transparencies of the desired subjects could not be obtained in anything like a systematic series, an assistant skilled in this branch of photography made negatives and slides from the specimens on exhibition in the public halls of the Museum of Natural History in New York, supplemented by copies of the best illustrations in standard works on natural history. There have been made some 800 negatives, in addition to a large number purchased from every available source. Such slides, although giving more satisfactory results when used in connection with the lime light, will be distinctly visible by from 50 to 75 persons when a lamp burning kerosene oil is used."



There's a lot worth saving in this country.

Today more Americans who value the best of yesterday are saving and using old buildings, waterfront areas and even neighborhoods.

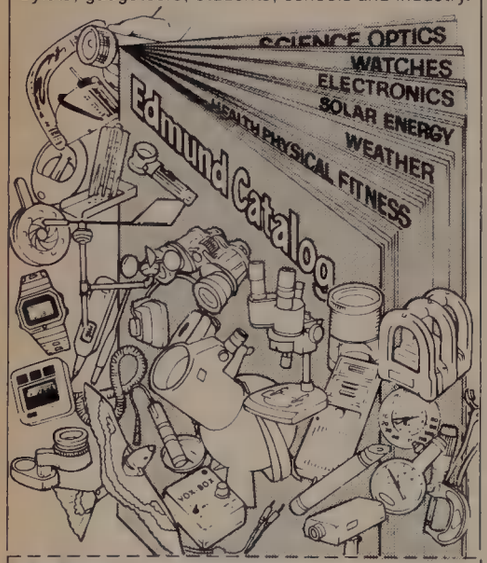
Preservation saves energy, materials and the artistry of these quality structures.

Help preserve what's worth saving in your community. Contact the National Trust, P.O. Box 2800, Washington, D.C. 20013.

National Trust for Historic Preservation
Preservation builds the nation

Send for Your Free Edmund Scientific Catalog Today!

America's largest mail-order market place of thousands of unusual and fascinating values for hobbyists, gadgeteers, students, schools and industry.



Edmund Scientific Co., Dept. R228
Edscorp Bldg., Barrington, NJ 08007
Please send me your FREE Catalog.

NAME

ADDRESS

CITY STATE ZIP

THE AUTHORS

PHILIP L. MARTIN ("Labor-intensive Agriculture") is associate professor of agricultural economics at the University of California at Davis. His interest in the subject of temporary alien workers originated when he was a Fulbright fellow in Germany in 1973 and 1974. He returned to the U.S. to obtain his Ph.D. in economics and agricultural economics from the University of Wisconsin in 1975. In 1978 and 1979 he was a Brookings fellow and also served as consultant on economics to the Select Commission on Immigration and Refugee Policy. He has been a member of several other commissions that were attempting to make policy on temporary workers both in Europe and in the U.S. Martin's chief interest is in the future of labor-intensive agriculture in postindustrial societies.

ROBERT W. CONN ("The Engineering of Magnetic Fusion Reactors") is professor of engineering and applied science at the University of California at Los Angeles. He got his B.S. at the Pratt Institute before going on to obtain his M.S. and his Ph.D. in engineering science from the California Institute of Technology. After getting his advanced degrees he spent a year at the Joint European Nuclear Research Center at Ispra in Italy and then joined the staff of the Brookhaven National Laboratory. In 1970 he joined the faculty at the University of Wisconsin at Madison and in 1974 he became director of the Fusion Research Program at Wisconsin, a job he held until moving to U.C.L.A. in 1980. Conn is codirector of the Center for Plasma Physics and Fusion Engineering there.

JOSEPH SILK, ALEXANDER S. SZALAY and YAKOV B. ZEL'DOVICH ("The Large-Scale Structure of the Universe") are astrophysicists who work at widely separated institutions. Silk is professor of astronomy at the University of California at Berkeley. His Ph.D. in astrophysics is from Harvard University. He has been at Berkeley since 1970, but he spent the past academic year (1982–83) as a visiting scholar at the Institut d'Astrophysique in Paris. He is the author of *The Big Bang* (W. H. Freeman and Company, 1980) and coauthor with John Barrow of *The Left Hand of Creation,* which is to be published by Basic Books later this year. Szalay is assistant professor of physics at Eötvös University in Budapest. He obtained his Ph.D. in astrophysics in 1975 from the same institution. After periods of postdoctoral study at Berkeley, the University of Chicago and Moscow University, he returned to Eötvös University in 1980. Szalay writes that his interest in cosmology was inspired by Zel'dovich, who is professor of astrophysics at Moscow University and director of the theoretical group at the Institute of Physical Problems in Moscow. Zel'dovich is a member of the Soviet Academy of Sciences and a foreign member of both the Royal Society in Britain and the National Academy of Sciences in the U.S.

JAMES E. DARNELL, JR. ("The Processing of RNA"), is Vincent Astor Professor at the Rockefeller University. He attended the University of Mississippi as an undergraduate, getting his B.A. in 1951. He obtained his M.D. at the Washington University School of Medicine in 1955. In 1956 he joined the staff of the Laboratory of Cell Biology at the National Institutes of Health. In that laboratory Harry Eagle had for the first time succeeded in culturing human cells outside the body. Together with Eagle, Darnell utilized the cultured human cells to do some of the first modern biochemical studies of the growth of viruses in their host cells. In 1961 Darnell moved to the Massachusetts Institute of Technology, and in 1964 he joined the faculty of the Albert Einstein College of Medicine in New York. He moved to Columbia University in 1968 and to Rockefeller in 1974.

ILANA STEINHORN and JOEL R. GAT ("The Dead Sea") are respectively an employee of Israel Limnological and Oceanographic Research, Ltd., and head of the isotope department at the Weizmann Institute of Science. Steinhorn, a native of Israel, received her B.S. in physics and mathematics (1972) and her M.S. in atmospheric sciences (1974) from the Hebrew University of Jerusalem. Her Ph.D. (1981) is from the Weizmann Institute. She writes: "My interest in the Dead Sea arose during my studies toward the Ph.D. degree. Because it was becoming apparent that the long-term stratification of the Dead Sea was weakening, the emphasis of my research was changed from a theoretical approach to one of intensive field measurements. These were set up quickly so as to document the changes in the structure of the Dead Sea water column that lead to its complete overturn." Gat was born in Germany and emigrated to the area then known as Palestine in 1936. He attended Hebrew University as an undergraduate and went on to obtain his M.S. in physical chemistry there in 1949. His Ph.D. in physical chemistry was awarded by the Weizmann Institute in 1956. After working for several years at the Israeli Atomic Energy Commission he

Enough already. No more inducements to business travelers, Qantas.



Before the coming of man, nobody had any business in the land Down Under. So, naturally, there was no business class.

Qantas changed all that. Qantas invented Business Class.

It gave rise to Business People who poured into Australia–until briefcases outnumbered koalas, wallabies and wombats. And word processors echoed across The Outback.

Not content with creating Business Class, Qantas improved it. Repeatedly. Now their extra-wide, deep-reclining seats recline even further. They've added fine china, linen and crystal service. Along with a roomy coat closet. And a complimentary flight bag.

Qantas Business Class now resembles First Class in other airlines. Still, the inducements continue:

Priority check-in and priority baggage handling. Special seating section with increased legroom.

Complimentary movies, complimentary stereo headphones (now they're padded!), and complimentary bar service.

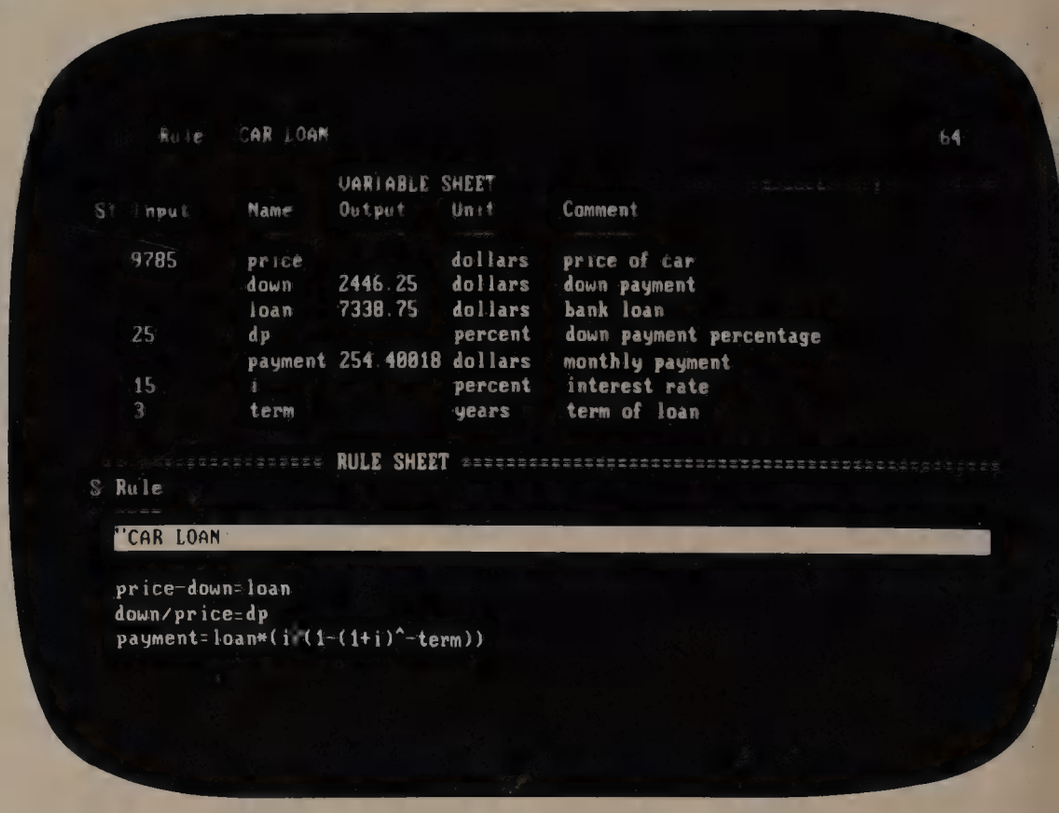
A choice of gourmet menu entrées, special wines and liqueurs. Special privileges at hotels, etc., etc., etc.

Qantas, we demand an end to such favoritism.

Koalas Against Qantas.

(The Australian Airline.)

TK!SO
EXPLA



TK!Solver does for equations what word processing did for words. The first thing you should know about the TK!Solver™ program is that it is not a spreadsheet. Instead, it does something completely unheard of (until now)—it turns your personal computer into a voracious equation processor.

The next thing you should know is that if the TK!Solver program can't make life with your personal computer easier (and pay for itself), even if you use it only 15 minutes a week, you are a very rare person.

And finally, you should know exactly what equation processing is, and how it works. If you keep reading this, you will.

Equation processing with TK!Solver, or problem solving made easy. The best way to understand what the TK!Solver program *is*, is to understand what it *does*. The following simple example is designed to do just that. If you're still a little in the dark after reading it, stop in at your local computer store for a very enlightening hands-on demonstration.

Begin by setting up your problem. The TK!Solver program lets you do it quickly, easily, and naturally. For example, a car costs $9785. What would be the monthly payment on a three-year loan if the down payment is 25% and the interest rate is 15%?

STEP 1. Formulate the necessary equations to solve your problem and enter them on the "Rule Sheet" simply by typing them in (as in the screen photo). For example: "price—down = loan."

STEP 2. Enter your known values the same way on the "Variable Sheet." For example: "9785" for price. You may also enter units and comments, if you want.*

STEP 3. Type the action command ("!" on your keyboard) to solve the problem.

STEP 4. TK!Solver displays the answer: the monthly payment is $254.40.

Backsolving, the heart of TK!Solver. Now that you've defined the problem and solved it, TK!Solver's unique backsolving ability also lets you think "backwards" to solve for any variable, regardless of its position in the equation. For example, if you can only afford a monthly payment of $200, you can re-solve the problem in terms of that constraint. The TK!Solver program will solve the problem, displaying your choice of a higher down payment, a longer loan term, or a lesser interest rate. This unique backsolving capability forms the basis of TK!Solver's remarkably flexible problem-solving ability.

OLVER™
AINED:

Also, as you can see from the example on the screen, TK!Solver deals not only with single variables, but with entire equations and sets of simultaneous equations. It also deals with much more complicated problems than this one. How complicated? That's up to you. What kinds of problems? That's up to you, too, but popular applications include finance, engineering, science, design, and education.

Other extremely useful and interesting things TK!Solver does. Aside from its basic problem-solving abilities, the TK!Solver program performs a number of pretty fancy tricks. Like: *Iterative Solving;* in which TK!Solver performs successive approximations of an answer when confronted with equations that cannot be solved directly, (like exp (x) = 2 – x • y and sin (x • y) = 3 – x – y). Like: *List Solving;* in which TK!Solver attacks complete lists of input values and solves them all, allowing you to examine numerous alternative solutions, and pick the one you like best. Like: *Tables and Graphs;* using the values you produced with the List Solver, the TK!Solver program will automatically produce tables and graphs of your data. You can look at your formatted output on the screen or send it to your printer with a single keystroke. And like: *Automatic Unit Conversion;* in which TK!Solver lets you formulate problems in one unit of measurement, and display answers in another. Very convenient what with all this talk about going metric.

The TK!Solver program also provides a wide variety of specialized business and mathematical functions like trig and log and net present value.

Then, there's TK!Solver's on-screen Help facility that provides information on commands and features any time you want it. Just type "?" and a topic name.

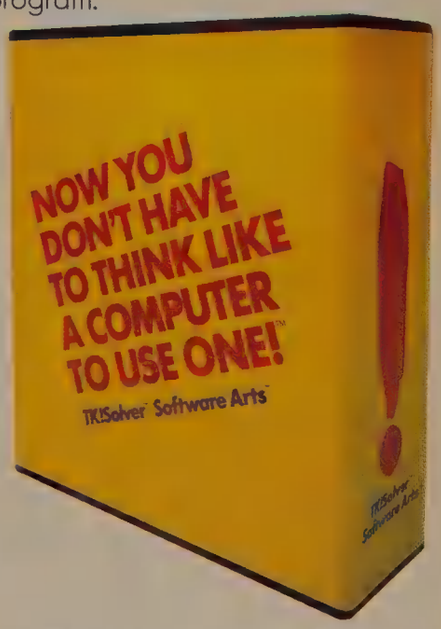
And of course the TK!Solver program combines all these features in one *integrated program.*

TK!SolverPacks make problem-solving a picnic. TK!SolverPack™ application packages are specially developed by experts in specific fields. Each package contains a diskette with about a dozen models that include the necessary equations, values, and tables for solving a particular problem. The models are usable as-is or you can easily modify them.

TK!SolverPack application packages available from Software Arts include Financial Management, Mechanical Engineering, Building Design and Construction, and Introductory Science. Additional TK!SolverPacks are on the way from Software Arts, McGraw-Hill,™ and others.

We know you're out there. No matter who you are, or what you do, if it involves using equations, the TK!Solver program is an indispensable tool for you.

So, visit your local computer store today, and see TK!Solver in action. You'll be amazed at how much faster and more effectively you'll be able to work when you discover the power of equation processing with the TK!Solver program.



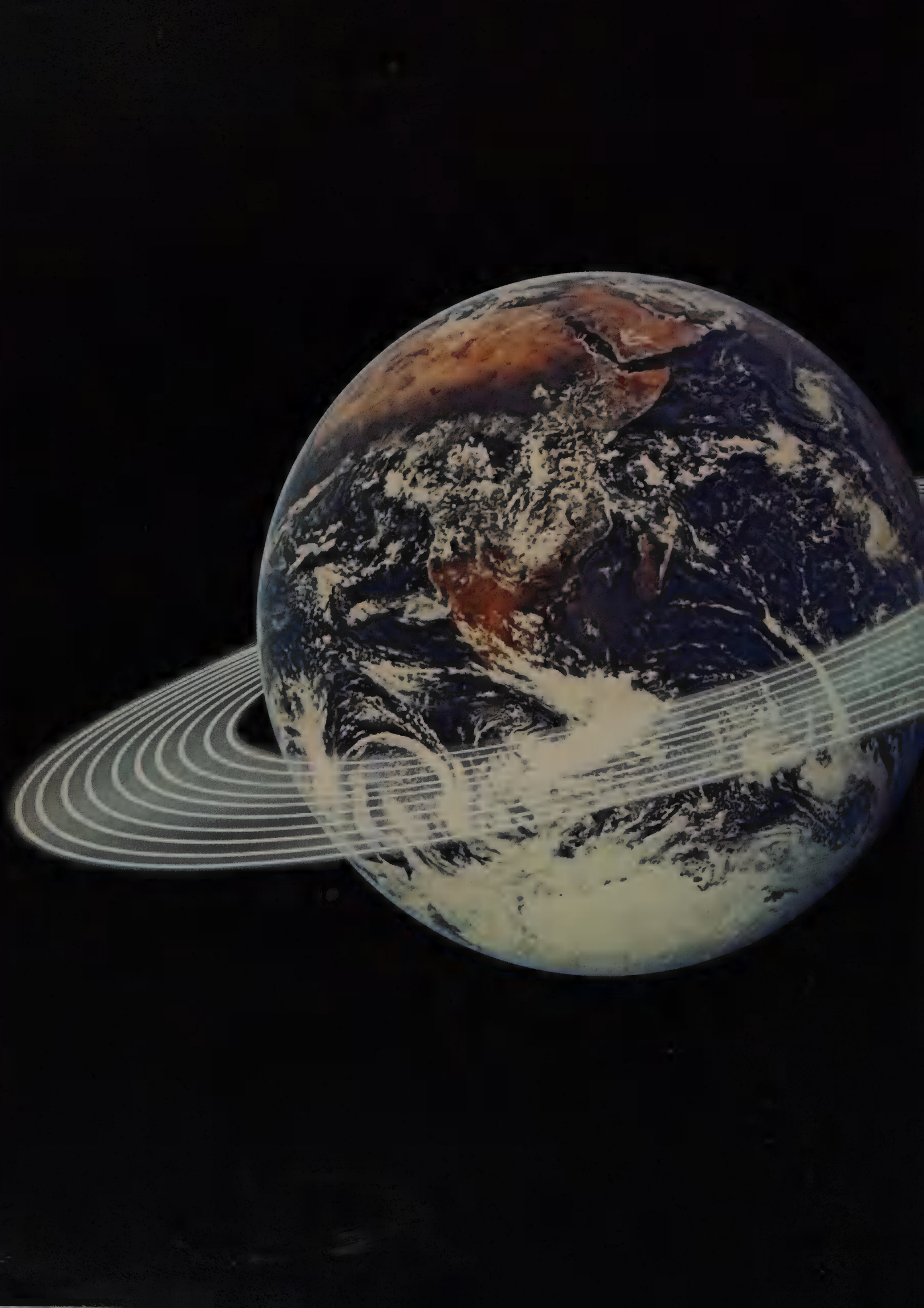
Software Arts™
The inventors of VisiCalc®

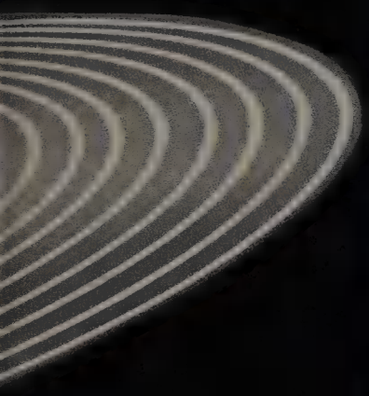
27 Mica Lane, Wellesley, Massachusetts 02181
617/237-4000

*You can easily define appropriate unit conversions on the unit sheet.

TK, TK!, TK!Solver, TK!SolverPack, The Problem Cruncher, the stylized ! and the slogan "NOW YOU DON'T HAVE TO THINK LIKE A COMPUTER TO USE ONE!" are trademarks or registered trademarks of Software Arts, Inc. SATN, TK!SATN and DIF are trademarks or registered trademarks of Software Arts Products Corp. Software Arts is a trademark of Software Arts, Inc. and Software Arts Products Corp. The TK!Solver program and the TK!SolverPack applications packages are products of Software Arts, Inc. which is solely responsible for their contents. VisiCalc is a registered trademark of VisiCorp.
Copyright © 1983 Software Arts, Inc. All rights reserved

P/N 100-092 P 8/83





Who on earth will install enough lightguide fiber to circle the earth 10 times?

Western Electric, an AT&T company.

In just five years, Western Electric has already created and installed enough lightguide fiber to circle the earth. That's 25,000 miles of fiber in 60 lightwave communications routes.

Those totals should increase at least tenfold by the year 2000, as Western Electric continues to install lightwave systems to transmit the high-speed voice and data services of the Information Age.

Far more efficient and economical than copper cable, lightwave communications speeds your messages over 90 million light pulses per second through hair-thin glass fibers. A single lightguide cable can carry 240,000 telephone calls at once.

Right now Western Electric is installing the world's largest lightwave system in the busy Northeast Corridor. Another will transmit the entire 1984 Los Angeles Summer Olympics to ABC television studios.

And our research and development arm, Bell Labs, is working on a transatlantic lightwave cable that will triple calling capacity and enhance transmission quality.

These systems are forerunners of a continually expanding lightwave network that one day will literally circle the earth, bringing everyone exciting new information services.

Western Electric, an AT&T company. Applying the technologies of lightwave, microelectronics, and software to make the dream of the Information Age a reality.

AT&T



QUESTAR® . . . *the telescope that can take it*

A Questar's durability is the thing that amazes so many owners. That an instrument so small and so beautiful can exhibit such toughness in the clinches is the quality that astonishes almost as much as its unique optical performance. Hubert Entrop, who sends us those great deep-sky photographs, keeps writing to tell us how he sometimes gets his best shots of galaxies and nebulas in areas that are exposed to sudden mountain squalls and gusting winds, but his Questar keeps guiding steadily along. Now comes a letter from *Wings*, a group that specializes in safaris to far away places. Will Russell, their leader, writes:

"WINGS now owns five Questars and during the last two years we've used them at Pt. Barrow, Alaska, in Kathmandu and Jerusalem, along the Amazon, in Kenya, Siberia, Thailand and many places in between. We've used them in the steaming tropics and sub-zero cold; in blizzards, dust storms, Atlantic hurricanes, Pacific typhoons and just about every other sort of weather. We've dropped them, kicked them over, watched them blow over, had them partially dismantled by foreign security guards, run over by airlines luggage conveyors and temporarily confiscated. In short, in the past 24 months we've learned quite a lot about using Questars."

And so, whether you propose to use your Questar on your patio or in one of those distant spots where Russell's Questars have been put through their paces, you will be reassured to know that Questars can take it. We like to think that everyone babies his or her Questar, but obviously Wings isn't able to do this and the instruments not only survive but turn in fine performances.

It is nice to know, too, that this old, established company still guarantees its product for ten years against defective parts and workmanship. Even Questars that have fallen down a flight of stairs or were dropped out second story windows, have turned up here to be put back together for a whole new useful life.

© 1982 Questar Corporation

Our booklet describing Questar, the world's finest, most versatile telescope, contains an astonishing collection of photographs by Questar owners. Send $2 for mailing costs on this continent; by air to S. America, $3.50; Europe and N. Africa, $4.00; elsewhere, $4.50.

QUESTAR

Box C, Dept. 202, New Hope, Pa. 18938

returned to the Weizmann as a faculty member in 1959.

B. O. K. REEVES ("Six Millenniums of Buffalo Kills") is associate professor of archaeology at the University of Calgary. He is a native of southern Alberta who was graduated from the University of Alberta with a B.S. in geology in 1961. He got a second undergraduate degree (a B.A.) two years later at the University of Alberta at Calgary before going on to obtain his M.A. (1967) in archaeology and his Ph.D. in the same subject (1971) at the University of Calgary. He writes: "My studies have been somewhat eclectic although geographically focused, dealing with archaeological sites ranging from those of early man to the period of World War II, including fur trading posts, missions, whiskey posts, and mine towns."

PETER WARD ("The Extinction of the Ammonites") is associate professor of geology at the University of California at Davis. He received his bachelor's and master's degrees from the University of Washington and completed his Ph.D. in 1976 at McMaster University. He taught earth history and paleobiology at Ohio State University from 1976 to 1978 before moving to Davis. He writes: "I am currently working on the chambered nautilus in New Caledonia, Fiji and Palau, using electronic methods to track the animals over several days and nights. The study of the living nautilus gives valuable information about the uses of the chambered shell and, more important, about the constraints imposed by the shell in the life cycle of the shelled cephalopods."

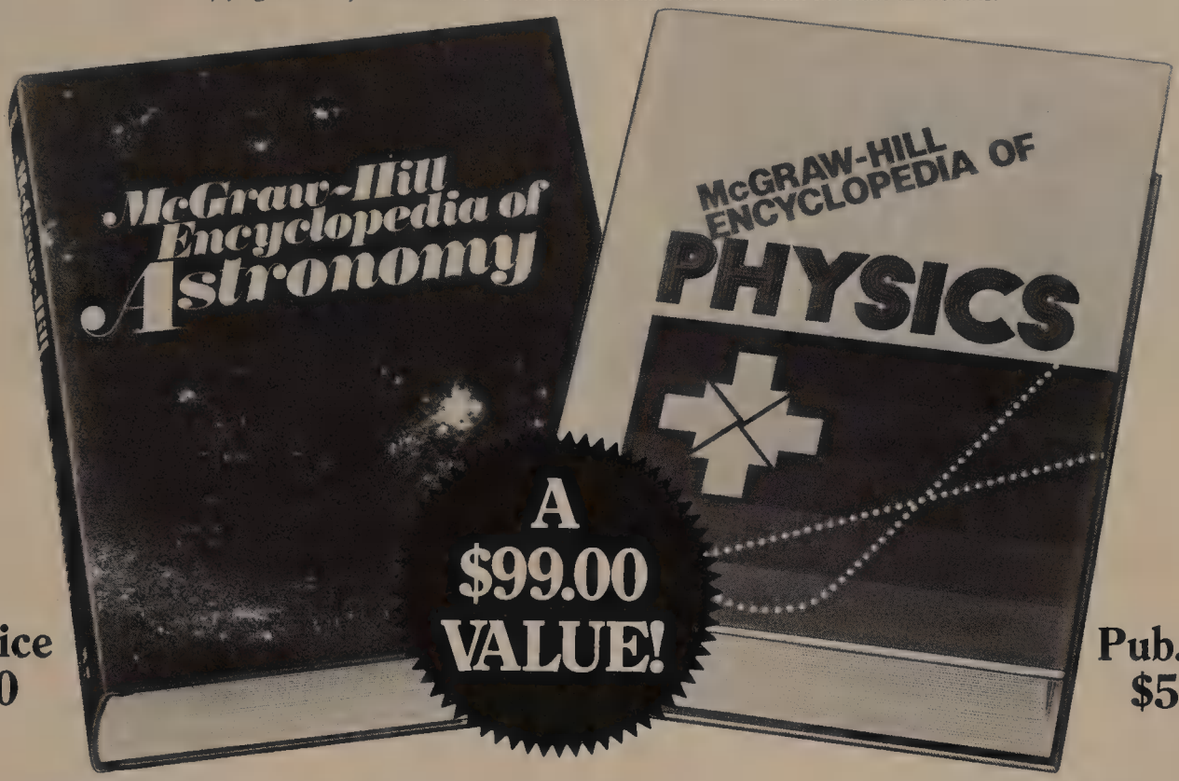
HERBERT F. YORK ("Bilateral Negotiations and the Arms Race") is professor of physics and director of the program in science, technology and public affairs at the University of California at San Diego. He got two degrees from the University of Rochester: an A.B. (1942) and an M.S. in physics (1943). His Ph.D. in physics was awarded in 1949 by the University of California at Berkeley. During World War II he worked for the Manhattan Engineer District (the "Manhattan Project") at the Y-12 plant in Oak Ridge, Tenn. In 1952 he helped found the Lawrence Livermore Laboratory of the University of California; later he served as its director. In 1958 he left to become the first Director of Defense Research and Engineering for the Federal Government, a job he was given by President Eisenhower. In 1961 he returned to the academic world as chancellor of the University of California at San Diego; he resigned in 1964 to become professor of physics. He has served as adviser to the Federal Government on arms and arms control in many different capacities.

An exhaustive compendium covering all aspects of physics and astronomy.

Take the 2-Volume LIBRARY OF PHYSICAL SCIENCES for only $3.95

as your introductory selection to one of Macmillan's two science book clubs—
The Astronomy Book Club *or* the Library of Science.

You simply agree to buy 3 more books—at handsome discounts—within the next 12 months.



Pub. Price $44.50

Pub. Price $54.50

The McGraw-Hill Encyclopedia of Astronomy contains 230 articles providing coverage of everything from constellations, and the gravitational lens to Einstein shift, and the instruments used in ultraviolet, radio and X-ray astronomy. Nearly 400 photographs, charts, graphs and drawings—including 13 pages in full color—illustrate and illuminate the articles.

The Astronomy Book Club is a very special club operated exclusively for astronomy enthusiasts. Whether you are interested in unusual astronomical happenings or just in being able to understand the night sky, whether you are an astronomer by profession or by interest, we're sure you'll find membership in the Astronomy Book Club beneficial.

Over 25,000 astronomy enthusiasts already are members of our Club. Recent offerings have ranged from popular explanations of deep-sky phenomena to authoritative references for serious sky watchers to important theoretical works on the frontiers of astrophysical research.

Eminent authorities such as Carl Sagan, Robert Jastrow, George Abell and John Gribbin appear frequently. The Club is devoted to making the finest books available, conveniently and inexpensively.

A SAMPLING OF THE INFORMATION YOU'LL FIND IN THESE VOLUMES:

Keeping you abreast of the latest advances in physics and astronomy, the Library of Physical Sciences provides a wealth of information on these closely related and rapidly developing fields. These two companion volumes contain a total of 990 articles by over 500 of the world's leading scientists. The articles originally appeared in the 15-volume *McGraw-Hill Encyclopedia of Science and Technology* (Fifth Edition, 1982).

These outstanding references provide easy access to material. Most articles contain a complete, current bibliography. Cross-references guide the reader to related topics. Finally, both volumes contain a detailed index.

MEMBERSHIP BENEFITS: In addition to getting the LIBRARY OF PHYSICAL SCIENCES for only $3.95 when you join, you keep saving substantially on the books you buy. Also, you will immediately become eligible to participate in our Bonus Book Plan, with savings of up to 70% off the publishers' prices. At 3–4 week intervals (16 times a year) you will receive the Book Club News, describing the coming Main Selection and Alternate Selections, together with a dated reply card. If you want the Main Selection, do nothing and it will be sent to you automatically. If you prefer another selection, or no book at all, simply indicate your choice on the card, and return it by the date specified. You will have at least 10 days to decide. If, because of late mail delivery of the News, you should receive a book you do not want, we guarantee return postage.

The McGraw-Hill Encyclopedia of Physics covers all areas of classical and modern physics as well as related topics in mathematics which play an important role in physical science. The 763 alphabetically arranged articles discuss experimental and theoretical aspects of acoustics, atomic and molecular physics, particle physics, fluid mechanics, heat and thermodynamics, low-temperature physics, nuclear physics, and more. You'll also find over 1000 tables, graphs, photographs and other illustrations.

The Library of Science is the oldest and most respected scientific book club in the country.

Convenient and enjoyable, The Library of Science puts you in touch with the major science developments, theories, and applications of our day. Books by Stephen Jay Gould, Karl Popper, and Steven Weinberg. Regularly made available to you in the Library of Science NEWS.

You'll save time selecting books that span an exciting variety of subjects from climatology to creativity, Einstein to evolving continents, with handsome discounts on virtually every book you purchase.

If reply card has been removed, please write: **Macmillan Science Book Clubs—Dept. ____-AZ3, Riverside, N.J. 08075** to obtain membership information and application.

10 QUESTIONS TO ASK YOURSELF BEFORE YOU BUY A BUSINESS COMPUTER.

The toughest thing about investing in the best computer is finding which computer is best for you. Because computers can be just as different as the people they work for.

That's why we've devised a list of 10 basic, unbiased questions to help simplify the choice of which computer will get along with you.

1 HOW MUCH SHOULD I SPEND?

As little as possible, and still get a fully functioning computer.

That would not include machines you see advertised for under $1,000. They're mainly designed for home entertainment and balancing your checkbook in full color.

For a serious business system, with all the capabilities you need, you can spend as little as $1,600, including software. Or as much as $5,000. And up.

Of course this gives you a choice of more than 500 different makes and models. So rather than devote the next 5 years of your life trying 500 different computers, a good short list would include computers from IBM, Kaypro, Apple and Tandy—the leading brands.

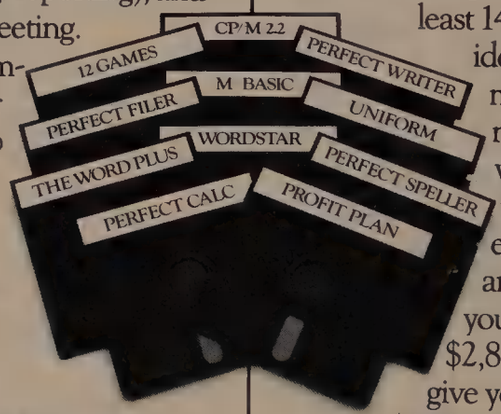
2 WHAT CAPABILITIES DO I NEED?

While businesses can be very different, their computer needs usually aren't.

We've found that 95% of all business needs can be fulfilled by a computer with three basic capabilities. Word Processing/Spelling, Data Base Management (filing/reporting), and Financial Spreadsheeting.

If the same computer has a popular 'operating system,' so much the better.

It will have thousands of other programs available to fill more specialized needs.



3 HOW MUCH HARDWARE DO I NEED TO GET STARTED?

Beyond the basic computer, you need a monitor and at least one disk drive, the device that stores information. For word processing or data base management, you will need a computer with 2 drives. Such powerful programs require the storage capacity of their own disk drive.

A second drive also eases copying data from one disk to the other.

4 HOW MUCH MEMORY WILL I NEED?

For most business purposes, a machine with a built-in memory (RAM) of 64K, about 42 typed pages.

Since you'll need a disk drive to store data, its capacity should be at least 140K. Of course, the idea is to get the most memory for the money. For example, when you compare the hardware of equivalent systems, an IBM PC will give you 320K for about $2,800. A Kaypro II will give you 400K for about $1,595. And an Apple IIe will give you 286K for about $2,400. So once again, the choice depends upon your needs.

5 DO I NEED AN 8-BIT OR 16-BIT SYSTEM?

A 16-bit system costs more, gives you a little more speed, and can run longer programs. Unfortunately, only a handful of programs

take advantage of 16-bit capacity.

75% of all microcomputers sold today are of the 8-bit variety, indicative of their ability to satisfy the needs of most businesses.

6 WILL I NEED A PRINTER?

Most people do, sooner or later. It's a fast, efficient way of getting information from your computer screen to other people. For letters, for reports, for balance sheets and for records. That's why it's important to make sure the computer you want has a built-in printer connection. Otherwise when you go to buy a printer, you'll also find yourself paying for an interface.

7 HOW MUCH TRAINING WILL I NEED TO USE MY COMPUTER?

You won't have to learn to write programs to use your computer. Somebody's already done that for you. You will have to learn how your machine operates, which may take a few hours.

And you'll learn how to run programs. The time that it takes depends on their complexity. The ideal way to save time is having a family of programs that share the same commands. Each program in the series will take less time to learn.

8 WHERE CAN I GO FOR GOOD SOLID ADVICE?

Do the same thing you would do if you were shopping for a stereo or an expensive camera. Ask an enthusiast. A computer buff. You probably know one. If not, call a consultant. In the long run, he or she could save you a lot of money by directing you to the right combination of hardware and software. And by advising you whether to buy a computer now or wait for prices to come down further.

It is true that the price of serious business computers is drifting down an average of 10-12% a year. The average system costing $3,000 today may cost $2,700 next year. Yet, you could lose thousands of dollars of increased productivity, waiting to save $300.

SPECIFICATIONS

Microprocessor
Z-80
Operating System
CP/M 2.2
User Memory
64K
Disk Drives:
2 drives, 400K, unformatted
Interfaces
1 Serial
1 Parallel
Keyboard
Detached, 63-key with numeric keypad
Software included:
Perfect Writer word processing
Perfect Speller
Perfect Filer
Perfect Calc spreadsheet
Wordstar word processing
The Word Plus
Profit Plan spreadsheet
M-Basic
12 Games
Uniform—allows computer to 'read' and 'write' TRS-80, Osborne, Xerox disks
Dimensions
Height: 8 inches
Width: 18 inches
Depth: 15½ inches
Weight: 26 lbs. (portable)

9 WHY IS THERE SUCH A GREAT DIFFERENCE IN THE PRICE OF COMPUTERS?

Some differences are based on capabilities. Business computers capable of playing joystick games cost more than ones that stick to business. And, as we mentioned before, 16-bit computers cost more than 8-bit systems.

A more important difference is the way computers are manufactured and sold. Most start with the basic keyboard and microprocessor, for a basic price.

You then pay more for a monitor. More for disk drives. More for communication interfaces. More for software. All optional extras that can run up the price two or three times.

However, at Kaypro we don't consider monitors, disk drives and software to be optional extras. So we make and sell completely integrated systems with all the hardware you need. All the software you need. All for $1,595.

Of course, a lot of other computer companies feel you should pay extra for everything you get.

10 WHICH IS THE SMARTEST COMPUTER FOR ME TO BUY?

All things considered, we'd consider it a Kaypro. And since this is a Kaypro ad, you wouldn't think we'd suggest anything else. But fortunately for us, our computer gives us a lot to be biased about. Off the shelf, it gives you everything you want for 95% of your business needs. Plus the ability to run 3,000 more programs. It's the complete business computer.

What do you get complete? 64K RAM, Z-80 microprocessor. CP/M, the most widely used business software operating systems. A 9" easy-to-read, green monitor. Two disk drives (400K capacity). Interfaces for both a printer and communications. A detachable keyboard, with a numeric keypad. And one of the leading computers in the $1,000 to $5,000 price range.

A Kaypro may not be the best computer for everyone. But for $1,595 complete, it gives you the best solution to serious business problems.

CALL 800-447-4700 FOR THE DEALER NEAREST YOU.

Registered Trademarks: Apple — Apple Computer, Inc., IBM — IBM Corp., CP/M — Digital Research, Inc., Z-80 — Zilog, M-Basic — Microsoft, Inc., Tandy, TRS-80 — Tandy Corporation, Osborne — Osborne Computer Corporation, Xerox — Xerox Corporation. Prices based on published information as of July 15, 1983. © 1983 Kaypro Corporation.



COMPUTER RECREATIONS

Introducing a department concerned with the pleasures of computation

by Brian Hayes

Let us calculate.
—GOTTFRIED WILHELM VON LEIBNIZ

Some of the livelier games played with a microcomputer now come equipped with a function designated TBIC. The letters stand for "The boss is coming," and when the key assigned to the function is pressed, the battlefield on the glass screen immediately goes silent and dark. Here, it seems, are the two poles of the public response to the recent proliferation of inexpensive computers. On the one hand the computer is an engine of business, a capitalist tool; on the other it is a medium for entertainments so frivolous that they must be hidden from view, like comic books.

I do not mean to belittle either the practical applications of computers in business and industry or the genre of computer games whose principal aim is to test the player's reflexes. Utilitarian computing is unquestionably important. As for the video games, their construction may well be among the highest expressions of the programmer's art. It should be observed, however, that neither of these uses of the computer engages very deeply the question of what a computer is and what it can do.

There is a vast territory between business programming and video games, between VisiCalc and Space Invaders. The territory includes the applications of the computer in all the arts and sciences, and perhaps most obviously in mathematics. It includes the use of the computer to simulate aspects of the natural world and of human societies. Furthermore, it includes many pursuits that properly speaking are not "uses" of the computer at all but rather serve to focus attention on the computer itself and on the nature of mechanized computation. It is this realm between stern practicality and mere diversion that "Computer Recreations" will undertake to explore. Contributions from readers are welcome. Given that thinking remains a good deal harder than computing, those without access to computing machinery should be at only a slight disadvantage.

Even the tools of the businessman can sometimes be applied to problems in the theory and practice of computing. Here I shall consider some questions raised by novel applications of the programs called electronic spreadsheets.

A paper spreadsheet is a large page ruled into many columns and rows. It might be employed for analyzing the budget of a company. Each department could be given a column and each category of income or expense a row. Totals and percentages for each department and each category could be entered in additional columns and rows.

The electronic spreadsheet reproduces this structure on the screen of a cathode-ray tube, but with a few notable differences. On paper a given cell (defined as the intersection of a column and a row) can hold either a label, such as the name of a department, or a number. In an electronic spreadsheet a cell can also be assigned a mathematical formula. Thus the cell at the end of a row might hold a formula that calls for summing the values entered into all the other cells in the row. What is displayed on the screen is the number that results from evaluating the formula, in this case the total, but the underlying content of the cell is the formula itself rather than the number. If one of the other entries in the row is changed, the total is recalculated automatically.

The first of the electronic-spreadsheet programs was VisiCalc, developed in 1978 by Daniel Bricklin, who was then a student at the Harvard University School of Business, Robert Frankston and Dan Fylstra. It is said to have sold more copies than any other computer program. Dozens of other programs operating on similar principles have been introduced since then, and VisiCalc itself has been revised several times. Most of the experiments described here were done with two later spreadsheet programs: 1–2–3, conceived by Mitchell Kapor and Jonathan Sachs of the Lotus Development Corporation of Cambridge, Mass., and Multiplan, a product of the Microsoft Corporation of Bellevue, Wash. In most cases other spreadsheet programs would serve as well.

Although the electronic spreadsheet was designed for financial analysis, it is capable of much more. It is a two-dimensional matrix of cells where the value of each cell can be made to depend on any other cell or group of cells. It is surprising how much of the mathematical structure of the world can be coaxed into such a format. Indeed, it turns out that the spreadsheet represents a quite general context for describing mathematical and logical relations.

A simple example can give a better grasp of how a spreadsheet is manipulated and suggest what its potential is. Each cell is specified by its coordinates in a grid; in most of the programs the columns are identified by letter and the rows by number, starting at the upper left. Suppose cells A1 and A2 are each assigned a numeric value of 1. In cell A3 a formula is then entered: the value at position A3 is set equal to the sum of the value in the cell immediately above it and the value in the cell above that one. In other words, A3 is equal to the contents of A2 plus the contents of A1, and it displays the value 2.

What has been accomplished so far is trivial: it is a highly elaborate scheme for expressing the relation $1 + 1 = 2$. It is now possible, however, to copy the formula in cell A3 into many other cells. (The exact procedure for making copies varies from one program to another, but all the programs include such a facility.) Suppose the formula in A3 is copied into cells A4 through A10. Each of those cells will then hold a value equal to the sum of the values in the two cells above it. Note that the formulas are all identical, but because they are applied to different values the results are not. The numbers displayed, reading from top to bottom, are 1, 1, 2, 3, 5, 8, 13, 21, 34 and 55.

There are many ways of generating the Fibonacci series with a computer, and most of them make far more efficient use of the machine's resources than this one does. There is something quite distinctive, however, about the spreadsheet strategy: it is nonalgorithmic. In almost all programming languages a task or the solution to a problem is defined in terms of an algorithm, that is, a sequence of explicit instructions to be executed one after another. An algorithm is like a recipe: it might begin, "First mix flour, yeast and water, then let rise and finally bake." Doing the same operations in another sequence would have a quite different result. The spreadsheet does not have this characteristic temporal ordering. What goes into the cells is not a sequence of steps that leads from the problem to the solution but a static structure that attempts

to encompass the entire procedure all at once. It is a description rather than a recipe: it states that bread consists of flour, yeast and water that have been mixed, allowed to rise and baked.

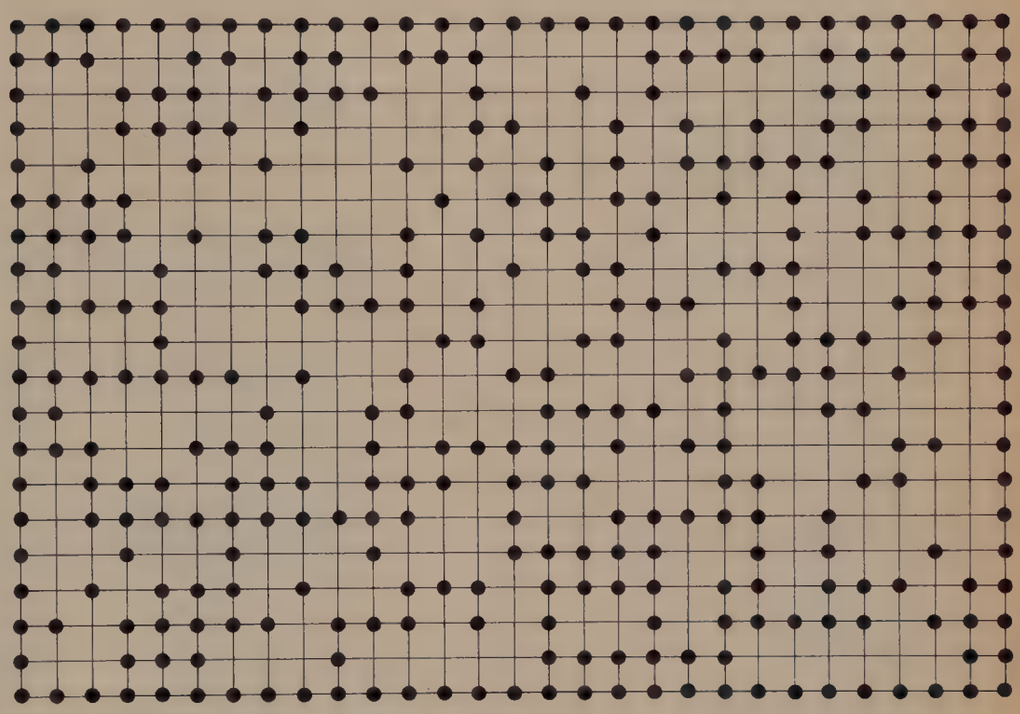
The distinction between an algorithm and a static description can be made clearer by another example. Consider the procedure for multiplying two matrixes of numbers when each matrix has three columns and three rows. The standard algorithm begins with instructions to multiply each element in the first column of the first matrix by each element in the first row of the second matrix, to add the three results and to store the sum as the first element of the product matrix. The same instructions are then repeated for the other eight combinations of rows and columns. The statement of the problem in a spreadsheet takes another form, and it can exploit the structural similarity of a mathematical matrix and an array of cells. Instead of writing a sequence of instructions, one simply defines the product matrix setting each cell equal to a formula that represents the appropriate combination of columns and rows. When the formulas have been entered, they are evaluated "all at once" and the entire product matrix appears.

At a deeper level, of course, a computer running under the direction of a spreadsheet program is indeed executing an algorithm. A computer that has only one central processing unit can do only one thing at a time, and so the cells are necessarily evaluated in some sequence. The user of the program, however, ordinarily has no need to take the sequence into account, and indeed he is often unaware of it. Thus the user need not think in terms of algorithms.

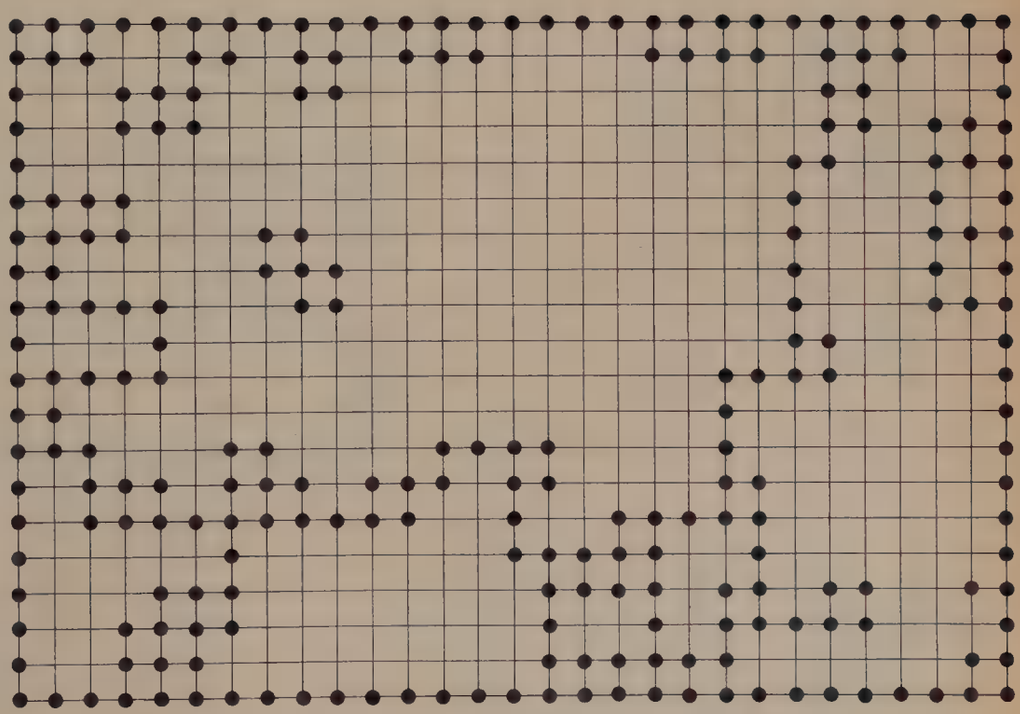
It is certainly not my intention to suggest that the nonalgorithmic mode of thought is somehow better than the algorithmic one. Some people may well prefer it, but that is largely a matter of taste. When a procedure becomes very complex, there is much to recommend the algorithm, which is more readily broken down into manageable pieces. Solving the problem all at once calls for understanding it all at once. It does seem likely, however, that there are certain problems or classes of problems that lend themselves naturally to a nonalgorithmic formulation.

One field where the spreadsheet's two-dimensional rectilinear array provides an appropriate medium is in constructing systems of cellular automata. The study of such systems was initiated in the 1950's by John von Neumann and Stanislaw Ulam, whose main concern at the time was with self-reproducing patterns. The rules they set themselves called for a "uniform cellular space" in which each cell represents an automaton, or machine, that has only a finite number of possible states. The space is uniform in the sense that the laws governing the state of the automata are the same for all the cells. A further constraint is that the state of a cell can be influenced only by its own history and by its close neighbors.

The conditions defining a system of cellular automata can readily be met by a spreadsheet program. In principle the number of states available to a cell is extremely large (perhaps 10^{100}), but it is clearly finite, and it can be reduced to a small number if that seems best; for example, a cell can be assigned a formula that can yield only two possible values, such as 0 and 1. The requirement of uniformity adds an interesting constraint. It implies that every cell in which a formula has been written must hold precisely the same formula. (There is more than one way of deciding whether two formulas are the same. Suppose a formula in cell A1 refers to cell B1, immediately below it. A formula in A2 might be considered identical if it also refers to B1, where the "absolute address" is the same, or if it refers to B2, where the geometric relation is preserved. The latter interpretation is generally more useful and seems more in keeping with von Neumann's and Ulam's ideas, but either scheme is acceptable if it is applied consistently. The spreadsheet routines for copying the contents of cells provide an easy operational test of uniformity. An array of cells can be considered uniform if a formula can be entered into one of



A lattice generated by a spreadsheet program simulates percolation



"Thinning" the lattice makes continuous paths easier to recognize

HAVE YOU HEARD?

The sound of hunger is louder than the rumble of an empty belly or the cry of a mother with nothing to feed her child. Hunger thunders through the generations and echoes against the dead end of abandoned dreams.

In the end, hunger can be heard in the scream of protest, in revolution and in rifle fire.

Oxfam America knows a better way. We work with people who are developing their own food and economic resources in 33 countries in Asia, Africa and Latin America. We put people to work in our own country, too, learning about the causes of hunger and what we, as responsible world citizens, can do about it.

We need your help to change the roar of hunger into a whisper of hope.

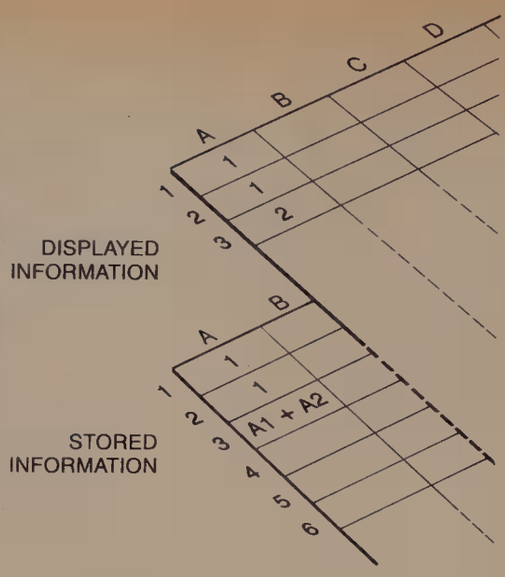


Oxfam America

Box N200
115 Broadway
Boston, Ma. 02116
617 482-1211

A full, audited financial report is available from Oxfam America, or, for New York residents, from the Department of State, Office of Charities Registration, Albany, NY 12231.





Structure of an electronic spreadsheet

them and then copied by the program into all the others.)

Von Neumann was able to prove that a self-replicating configuration of cells does exist. He did it the hard way, by showing there is a universal constructor that can create any pattern and therefore must be able to create its own pattern. The proof calls for some 200,000 cells having 29 possible states. As far as I know, the pattern has never actually been constructed, either manually or with the aid of a computer. Conceivably it might be within the capabilities of the larger spreadsheet programs.

A far simpler system capable of self-replication was devised in 1960 by Edward Fredkin of the Massachusetts Institute of Technology. Each cell has just two possible states, living and dead, which can be represented by the numbers 1 and 0. The state a cell will occupy in the next generation is determined by the present state of its four orthogonally adjacent neighbors, that is, the four cells immediately to the north, east, south and west. If the number of living neighbors is even (0, 2 or 4), the cell dies or remains dead. A cell with an odd number of living neighbors (1 or 3) lives.

It is a straightforward matter to express this rule in a spreadsheet formula, particularly with programs that include a function for modular arithmetic. In the case of cell B2 the formula is (B1 + C2 + B3 + A2) modulo 2. The effect of the formula is to add the values in the four neighboring cells, divide by 2 and retain only the remainder, which is necessarily either 0 or 1. It then remains only to copy the formula (in such a way that cell references preserve the same geometric relations) into all the cells throughout a region of the spreadsheet. Actually there is one further subtlety in the construction of the system: two copies of the cellular space are needed. One copy represents the present generation and one preserves the state of the preceding generation. The present state of a cell is based on the number of neighbors it had in the earlier generation.

When the spreadsheet is set up according to Fredkin's rules and an initial pattern is supplied, each cycle of recalculation yields a new pattern. After a few cycles four copies of the original configuration appear. Later the copies themselves are copied four times, so that the initial pattern has been reproduced 16 times. The number of cycles needed for reproduction depends on the complexity of the initial pattern; in the simplest case (a single live cell) the four offspring appear immediately.

Watching the progress of a growing colony can be fascinating. Fourfold symmetry is maintained at all times, and some of the patterns have a striking, stellate form. There is a rhythm to the process: the perimeter of the occupied area expands continuously, but the interior periodically becomes filled with a dense thicket of cells and then empties again.

Surely the best-known system of cellular automata is the game of life, invented by John Horton Conway of the University of Cambridge and introduced to the world at large in 1970 by Martin Gardner in his *Scientific American* column "Mathematical Games." The game, like its own pullulating automata, has by now spread to virtually every kind of computer system and programming language. There is a good reason: the game rewards the attention given it, whether the attention takes the form of casual spectatorship or close analysis.

In Conway's game of life the rules are defined not to ensure the replication of a pattern but rather to maximize variety or minimize predictability. Each cell again has two possible states, but the surrounding neighborhood is made up of the eight nearest cells, including the diagonally adjacent ones. If a cell is living, it will continue to live in the next generation only if it has either two or three living neighbors. With fewer neighbors it is said to die of loneliness and with more it dies of overcrowding. For a nonliving cell a birth is possible only if there are exactly three living neighbors.

An algorithmic specification of this procedure tends to be highly repetitious. It calls for examining a given cell, counting its neighbors, deciding whether the cell is to live or die, then going on to the next cell and the next until all of them have been checked. The repetition is generally embodied in the program structure called a loop, which is executed once for each cell. When the game of life is encoded in a spreadsheet, the repetition is still there, but it is spatial rather than temporal: the same formula is entered into each cell in a large array.

ITT announces the biggest advance in television since color.

Your everyday television set is about to become computer-like.

And it will be doing some very un-television-like things.

It's all due to the item you see here—a sophisticated set of eight semiconductor chips developed by the people of ITT.

These tiny chips will replace some 350 parts in today's television set.

Using the latest in digital technology ("digital" being the language of computers), they will be able to provide astonishingly clear pictures (clearer than anything *any* TV had before). As well as superb, new sound quality.

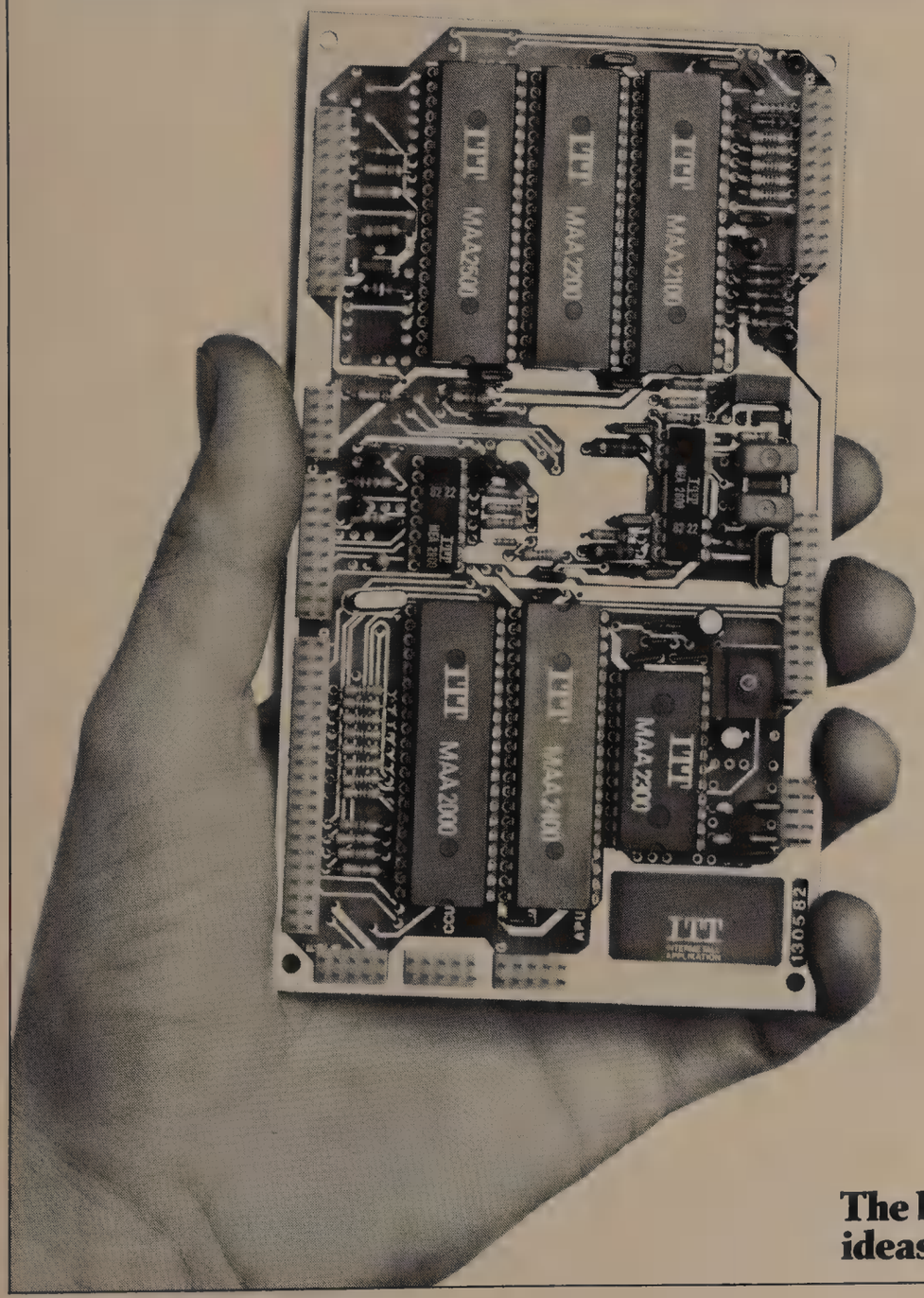
You'll also be able to have two channels on your TV screen at once. Finding out how your stock did on one channel, for example, while you watch the late news on another.

With this new kind of TV, it will be easy to shop at home. You'll just punch up your purchase on a keyboard, selecting from a catalog on the TV screen.

Your TV set will even become part of a home security network—allowing you to have surveillance over your whole property at the touch of a finger.

ITT is making these chips available to various television manufacturers, who will be marketing their own "digital" TV sets very soon.

It's a television worth watching for.



The best ideas are the ideas that help people. ITT

© 1983 ITT, 320 Park Avenue, New York, N.Y. 10022, U.S.A.

Teach yourself speed reading in 10 hours or your money back!

1-800-228-5655

Most executives must read a million words each week if they are to do their job.

1,000,000 words! And that's reading related to immediate business. It does not include other business related reading—you know the reading that helps you stay ahead of the game.

That's why most executives spend almost half of every work day reading. About 3 hours a day is average.

The good news is that now you can cut this load in half—at the very least.

You can use Speed Reading Self-Taught (SRST) to teach yourself speed reading. In just 10 hours or so you can double or triple your reading and increase your comprehension. Xerox Learning Systems guarantees this or your money back.

Speed Reading Self-Taught was originally created especially for our clients. (We help thousands of companies help their people do a better job, including 357 of the Fortune 500).

Thousands of executives in sales, data processing, engineering, marketing and management have proved out the SRST learning system, reporting significant gains in reading speed as well as impressive lifts in comprehension. SRST works for slow readers...for fast readers. It will work for you. The materials are mailed to you complete. All you need is a cassette recorder and a watch with a second hand.

SRST costs $95. Your organization will probably pay or reimburse you. We guarantee SRST will teach you lasting speed reading skills or your money back.

Most executives read at a seventh grade pace. Isn't it time you broke out of that pack?

You can teach yourself speed-reading in 10 hours with SRST. You can double or triple your reading speed and increase your comprehension. Xerox Learning Systems guarantees this or your money back.

To order Speed Reading Self-Taught, simply call us, toll free, at 1-800-228-5655. In Nebraska, 1-800-642-8777. Major credit cards accepted or send check for $95.00 plus $4.95 shipping and handling to Xerox Learning Systems, 6 Commercial Street, P.O. Box 944, Department SRST-S9ASA, Hicksville, N.Y. 11802.

XEROX

Xerox Learning Systems

There are many ways of writing a formula to evaluate the state of a cell in the game of life. The best I have seen (meaning the one that runs the fastest) was devised by Ezra Gottheil of Lotus. The basic procedure is to multiply the value in the cell under examination by 9, yielding a result of either 0 or 9, then add the values in the eight surrounding cells. That result is looked up in a small table that gives the new state of the cell for all possible values of the sum (namely those between 0 and 17).

Ideally the game of life would be played on a cellular matrix of infinite extent. One of the fascinations of the game is that certain small initial patterns grow into magnificent symmetrical blooms after just a few generations; other patterns emit compact projectiles that glide off into the indefinite distance. The evolution of the pattern is changed whenever an organism falls off the edge of the world. An infinite matrix is impossible under any circumstances, and when one is working with a spreadsheet, the practical limits are indeed rather small. They are set by the capacity of the program itself, by the memory capacity of the computer and by one's own patience. (The time needed for creating a new generation is roughly proportional to the number of cells included.) One strategy for creating an array that has no boundaries even though its area is finite is to define the cells along opposite edges as being adjacent; the effect is to change the topology of the spreadsheet. Joining two edges in this way creates a cylinder or, if the sheet is twisted, a Möbius band. Joining all four edges in pairs forms a torus.

The Ising model is a simulated physical system that superficially resembles some cellular automata, although its interpretation is quite different. The model, which was introduced in the 1920's by the German physicists Wilhelm Lenz and Ernest Ising, can represent a number of physical phenomena, but it is most commonly applied to the description of ferromagnetic materials. Each site in a lattice represents the spin angular momentum, and hence the magnetic moment, of an atom. Each spin has a fixed magnitude, but the spin axis can point either up or down. When all the spins point in the same direction, the material is fully magnetized; when the spins are random, the magnetization is zero.

A spreadsheet realization of the Ising model is somewhat more complicated than one for Fredkin's replicating automata or Conway's game of life. The state of a given spin is again influenced by the nearest-neighbor cells, in this case the four orthogonal ones. In the Ising model, however, there is an element of randomness, which represents the effect of nonzero temperature. If a cell's neigh-

Radio Shack's Pocket Computers —Have One To Go!

TRS-80® PC-2
199.95
26-3601



TRS-80 PC-4
69.95
26-3650

TRS-80 PC-3
99.95
26-3590

TRS-80 PC-2. A hand-held computer with power and speed that rivals many desktop micros, yet measures just $1^{1}/_{16} \times 7^{11}/_{16} \times 3^{3}/_{8}$". Ideal for engineers, business people, scientists— anyone with tough problems to solve on the go! Programmable in Extended Pocket BASIC. Fully-addressable 7 x 156 dot-matrix LCD for graphics, plus upper and lower case characters. Expandable 2640-byte memory. Add the PC-2 Printer/Plotter/Dual Cassette Interface (26-3605, $219.95) to create spectacular four-color charts and graphs or to use our library of ready-to-run software.

TRS-80 PC-3. Perfect for students. The PC-3 is easily programmable in Pocket BASIC, and features 16 arithmetic and 8 string functions, 1.4K memory, 24-character display, two-digit exponents, multiple statements, arrays and more. $^{3}/_{8} \times 5^{5}/_{16} \times 2^{3}/_{4}$". Add our Printer/Cassette Interface (26-3591, $119.95) and choose from our large library of programs.

TRS-80 PC-4. Our lowest-priced pocket computer ever! Keep up to ten short programs in memory and enter BASIC commands with just two keystrokes. Features expandable 544-character memory and 12-character display (scrolls to 62 characters). Just $^{3}/_{8} \times 6^{1}/_{2} \times 2^{3}/_{4}$". Expand with the Cassette Interface (26-3651, $39.95) and Printer (26-3652, $79.95).

TRS-80 Computers Are Available Nationwide. Drop by any Radio Shack Computer Center, participating store or dealer and pick up your own TRS-80 Pocket Computer today.

Radio Shack®
The biggest name in little computers®
A DIVISION OF TANDY CORPORATION

☐ **Send me your free TRS-80 Computer Catalog today!**

Mail To: Radio Shack, Dept. 84-A-701
300 One Tandy Center, Fort Worth, Texas 76102

NAME _____

ADDRESS _____

CITY _____ STATE _____ ZIP _____

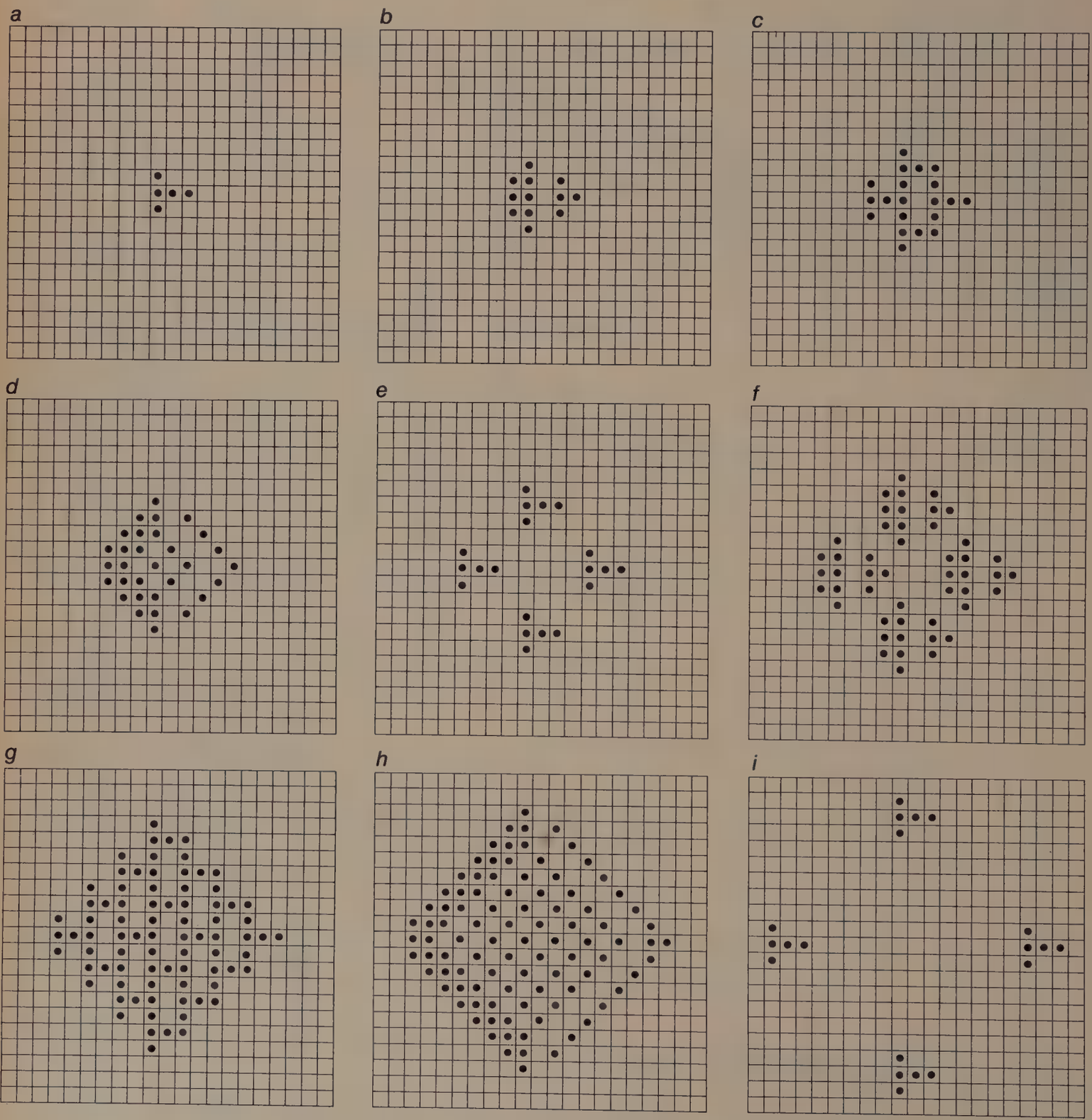
TELEPHONE _____

Prices apply at participating Radio Shack stores and dealers.

bors are all pointing up, the cell also has a tendency to point up, but it is not certain to do so; the probability is inversely proportional to the temperature.

A few experiments I have undertaken with a spreadsheet Ising model have yielded mixed results. The expected properties of the two-dimensional model have been known exactly since 1944, when Lars Onsager of Yale University solved the system analytically (rather than by numerical simulation). As the temperature is lowered through a threshold value (the Curie temperature) the spins should begin to fluctuate wildly and then should become fully magnetized. I have not seen such a phase transition, but I am not really surprised. The Ising model makes outrageous demands on computational resources. Getting accurate results requires a large lattice and a consideration of all possible configurations of the spins, which can take many hours even with an efficient program and a high-speed processor. At the Institute for Theoretical Physics of the University of California at Santa Barbara a special-purpose Ising-model computer has been built that calculates 25 million spins per second. The corresponding figure for the spreadsheet version of the model is about 25.

Even though the interesting events near the Curie temperature cannot be observed, the spreadsheet Ising model does seem to embody some other properties of magnetic materials. At high temperature the array of spins is without apparent pattern, as one would expect. At low temperature the magnetization of the lattice is obvious. Large, irregular blocks of aligned spins develop spontaneously, and blocks of opposite polarity seem to contend along their boundaries. One surprise (at least it surprised me) was the appearance of an antiferromagnetic phase, in which every other spin points in the opposite direc-



System of cellular automata devised by Edward Fredkin in which any pattern reproduces itself

Win a trip to Scotland.



Enter the J&B "It Whispers" Photo Contest.

We're looking for the photograph that whispers J&B best. Capture that scene, and J&B will cover all expenses for a once-in-a-lifetime trip to Scotland. Plus, for the best of the rest, we've got 160 other fabulous prizes.

Grand Prize

Ten-day, nine-night trip for two to Scotland, including round trip air transportation, hotel accommodations, meals, car rental, visit to J&B distillery, sightseeing, plus $2,500 spending money.

10 First Prizes

Nikon F3 Camera with 50mm F1.8 Nikkor Lens. 35mm Camera features automatic exposure control, compact and light design with built-in body grip. Uses Nikon Bayonette Mounting system which assures fast and accurate mounting.



50 Second Prizes

Nimslo 3-Dimensional Camera Outfit. Fully automatic, uses popular brands of 35mm color print film. View finder and information display. Small compact styling and light-weight construction.



100 Third Prizes

Vivitar Instant Slide Printer. Makes Polaroid Instant 3¼" x 4¼" print from any 35mm slide. It features automatic exposure control, a variable composition selector and a built-in electronic flash.



How to Enter

Use the coupon on this page or see your participating liquor store for an entry blank. Take a picture of the scene you think whispers J&B, and send it in. Be sure to follow the rules below. Incomplete entries will be invalid.

Official Rules— No Purchase Necessary

1. To enter, submit a black-and-white or color photograph (2¼" x 2¼" to 8" x 10") of a scene that you feel "whispers." Slides and transparencies are not accepted.
2. Handprint your name, address, and zip code on the official entry form or a plain piece of paper. Glue or tape it to the back of your photo entry. Do not write on photo. Include cardboard backing to protect picture in mail. Include with your entry the answer to the following question: What do the initials "J&B" on a bottle of J&B Rare Scotch stand for? Your entry will not be eligible for judging unless this question is answered. The information needed to answer this question may be found by looking at the label of any bottle of J&B Rare Scotch. You may obtain a free label facsimile by writing to: J&B Label, P.O. Box 3244, Syosset, NY 11775.
3. Enter as often as you wish, but each entry must be mailed separately to: J&B Third Annual "It whispers" Photo Contest, P.O. Box 3258, Syosset, NY 11775. Entries must be received by December 31, 1983. No responsibility is assumed for lost, misdirected or late mail.
4. Entries will be judged on the basis of originality (0 to 30 points), relevance to the theme (0 to 40 points), composition (0 to 20 points), photographic technique (0 to 10 points). Winners will be selected under the supervision of National Judging Institute, Inc., an independent judging organization whose decisions are final on all matters relating to this contest. All prizes will be awarded and winners notified by mail. Prizes are not transferable or exchangeable. Only one prize to an individual or family.
5. All entries become the property of The Paddington Corporation with all rights, including the right to edit, publish and use any photo without further consideration of payment to the contestant. No correspondence about entries will be entered into, nor will photos be acknowledged or returned.
6. Before receiving a prize, each entrant must warrant his age, that the photograph was taken by himself, that he is an amateur photographer, and that he has full rights to the photograph and that it has won no previous award or competition.
7. The contest is open to U.S. residents, except employees and their families of The Paddington Corporation, its affiliates, advertising and sales promotion agencies, liquor wholesalers and retailers, and professional photographers. Void where prohibited. All federal, state and local regulations apply. Taxes on prizes, if any, are the responsibility of the individual winners.
8. Entrants must be of legal drinking age in the state of their residence as of September 1, 1983. For a list of major winners, send a stamped, self-addressed envelope to: J&B Third Annual "It whispers" Photo Contest Winners, P.O. Box 3269, Syosset, NY 11775. Winners list will be available as of March 15, 1984.



Official Entry Form

Mail your completed entry & photograph to: J&B 3rd Annual "It Whispers" Photo Contest, P.O. Box 3258, Syosset, NY 11775

Name_____

Address_____

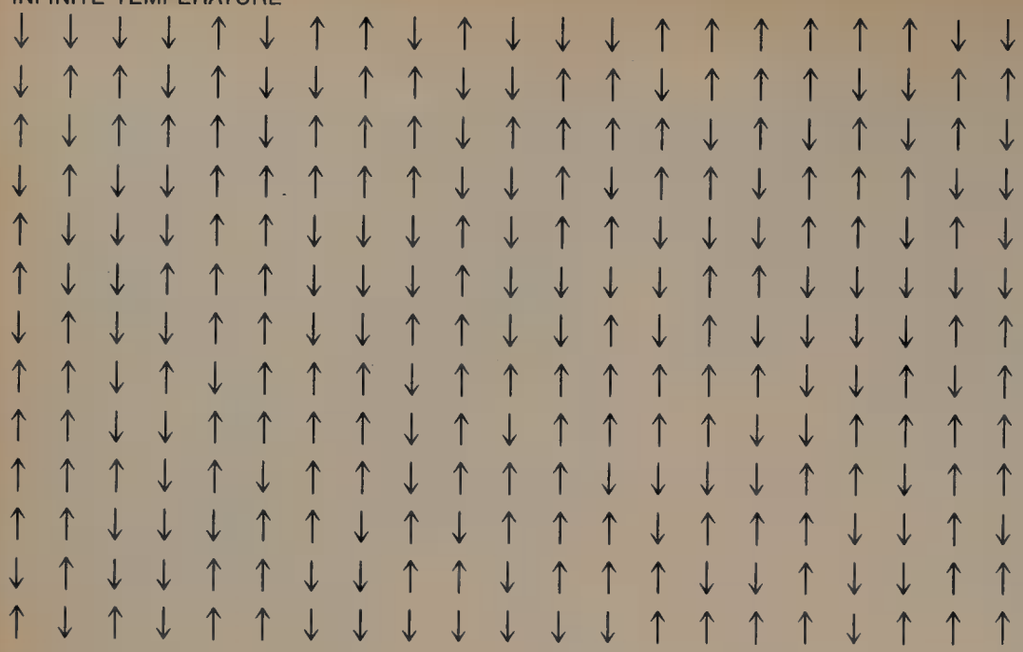
City_____ State____ Zip____

The initials on the label of a bottle of J&B Rare Scotch stand for:_____

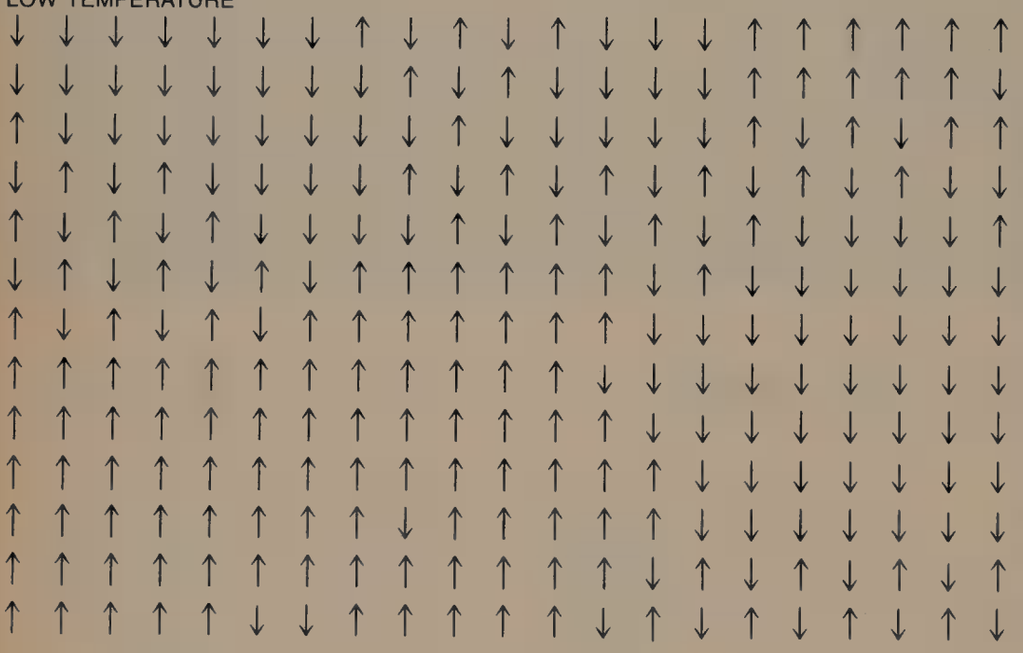
Void where prohibited. No purchase necessary. SAM

86 Proof Blended Scotch Whisky, © 1983 The Paddington Corp., NY

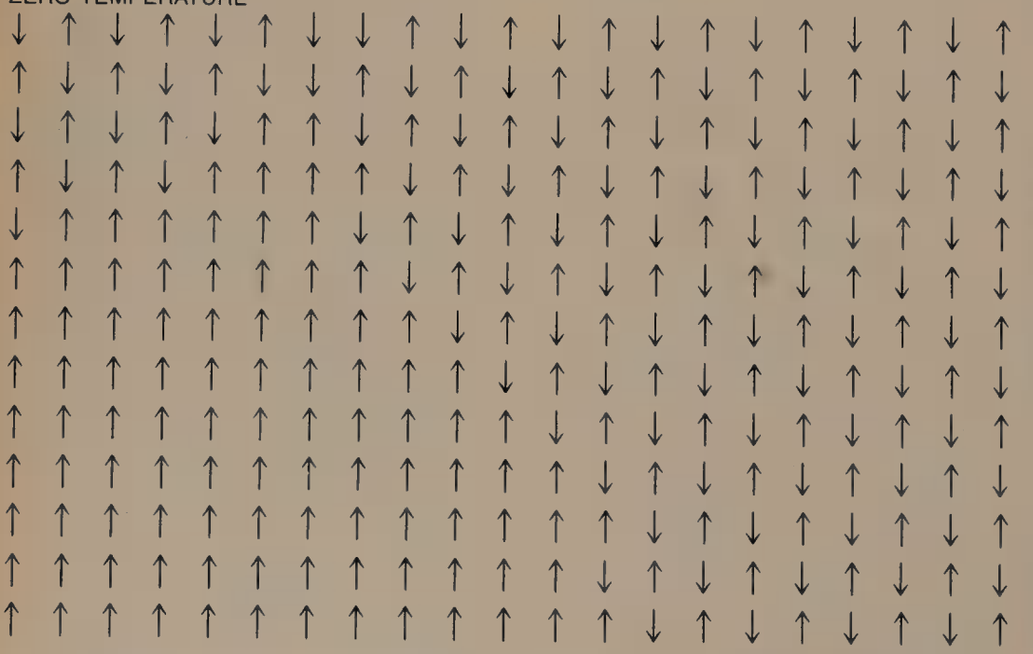
INFINITE TEMPERATURE



LOW TEMPERATURE



ZERO TEMPERATURE



A spreadsheet version of the two-dimensional Ising model of ferromagnetism

tion. At a temperature of absolute zero the antiferromagnetic phase seemed to be the stablest configuration, but this may reflect a deficiency or an error in my implementation of the model.

Other physical systems can also be conveniently dealt with on a rectilinear lattice. One phenomenon that interests me is percolation, which describes not only the brewing of coffee but also the structure of some polymers, the conductivity of alloys, the efficiency of telephone networks and the propagation of forest fires and infectious diseases. In one simple percolation model the cells of a lattice might represent possible sites of copper atoms in an insulating matrix. As a first approximation, the probability that a site is occupied is proportional to the bulk concentration of copper. The quantity of interest is the probability of forming an unbroken chain of copper atoms across some domain of the lattice; it is this probability that determines the conductivity of the material.

A percolation model in a spreadsheet again demands some form of random function. The easiest approach employs a uniform cellular space, where all the sites have the same probability of being occupied, and they are all independent of one another. The result is a random array of filled and vacant cells. It is then necessary to determine whether or not there is a continuous path across the space. A proper solution would be to mount a systematic search of each potential path, but it is not clear to me how that can be done without resorting to algorithmic methods. A cruder but still helpful technique is to thin out the underbrush by including in the formula for each cell a function that eliminates any atom with fewer than two neighbors. Because any atom that forms part of a chain must have at least two neighbors, the chain is not affected by this procedure, but many dead ends and isolated clusters wither away with each recalculation of the spreadsheet.

A spreadsheet program is surely not the ideal medium in which to represent any of these mathematical or physical models. For serious work each of them would have to be embodied in its own special-purpose program. In the case of the game of life I have compared various spreadsheet versions with a program (an algorithm!) written in the native language of a microprocessor. For arrays of the same size the machine-language program is almost 100 times faster than the fastest spreadsheet. The reason is not hard to find: even though a cell can never have a value other than 0 or 1, the spreadsheet program calculates the value to 15 decimal places.

If the spreadsheet cannot claim efficiency, however, it has the compensating virtue of versatility. Writing a machine-language program for playing the

TECHNICIANS, MANAGERS, INVESTORS

SUBSCRIBE TO High Technology Magazine AND PROFIT FROM OTHER PEOPLE'S SUCCESS.



Tooker
High Technology
August 1983

"If microprocessor chips are the engines of this information age, then memory chips are the fuel which will be consumed in ever expanding quantities; memories are a strategic business for us."

Gary L. Tooker
Sr. VP & General Manager
Semiconductor Products Sector
Motorola



Jones
High Technology
July 1983

"Flexible manufacturing is expanding and will account for a significant segment of total machine tool sales. It also yields an excellent return on investment."

Frank W. Jones, President
Giddings & Lewis
(unit of AMCA International)



Hall
High Technology
August 1983

"45 percent of productivity gains in the United States over the past six years was attributable to technology . . . and one-third of that came from computers. The next step must come from improved communications."

Robert C. Hall, President & CEO
Satellite Business Systems



Simon
High Technology
March 1983

"Our clinical laboratory analytical systems are based on traditional technologies. But in the next ten years, the industry will introduce new technologies that are different in kind from anything available today. Major advances in immunology will radically alter the practice of laboratory medicine."

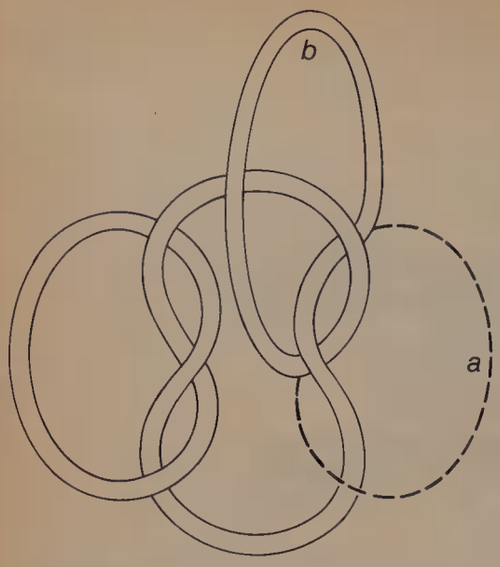
Henry Simon, President
Technicon Corp.

Whether you're an engineer, technician, manager, or investor—or a combination of all three—you can reap the rewards of the technological revolution. Let High Technology magazine show you how. With articles that acquaint you with advances in fields outside your own. Inform you of important trends emerging in companies like yours. And introduce you to the people and corporations that are making things happen today in the industries of the future. As high technology transforms the world we live in, rare opportunities for success are becoming unlimited opportunities overnight. See for yourself. Take this opportunity to try a free issue. There's no obligation to subscribe.



Please use the adjacent subscription card to order your free copy plus 11 additional issues, for just $21.00 ($1.75/copy), or write to:

highTechnology
P.O. Box 2808, Boulder, Colorado 80321



Proof that a square knot is alternating

game of life is more than an evening's entertainment. Moreover, the program can do nothing else, whereas the simple matrix of linked cells in a spreadsheet constitutes a problem-solving device of impressive generality. There is much more, beyond the ideas sketched above, that is clearly within the capabilities of the programs. It appears that any series of numbers in which the terms are defined by algebraic or trigonometric functions can be generated. A prime-number sieve can be constructed out of one short formula, repeated some hundreds of times. A physical field can be represented by allowing the address of each cell to serve as its coordinates in two-dimensional space. As the sales brochure invariably says: "The only limit is your own imagination."

Is it true? Can a matrix of interdependent formulas without an algorithmic structure be made to compute anything that is computable? Is the mechanism not merely general but universal? In the case of an infinite matrix the question has been settled. Conway has proved that the cellular world of the game of life has sufficient resources for the construction of a Turing machine, the conceptual model of a universal computer. Since an infinite spreadsheet could be employed to play the game of life, it could also be employed to create the Turing machine.

Such a result is certainly worth knowing, but even if the requirement of infinite area could be relaxed, the demonstration would be of no practical significance. Life is too short, and the game of life too long. There is a less formal approach to measuring the scope of the spreadsheet programs that I find more promising. It is the hit-or-miss method of applying the programs to various problems and exercises from the stock of old favorites in computer science. The interesting test cases are likely to be the ones with a highly efficient algorithmic solution. One example is the Tower of Hanoi puzzle, in which several rings are stacked in order of decreasing size on one of three pegs; the aim is to move the rings one at a time, without ever allowing a larger ring to rest on a smaller one, until they are stacked in the same order on another peg. The standard solution employs a recursive algorithm, one that states the final stage in the procedure explicitly and then calls on itself to define the earlier stages.

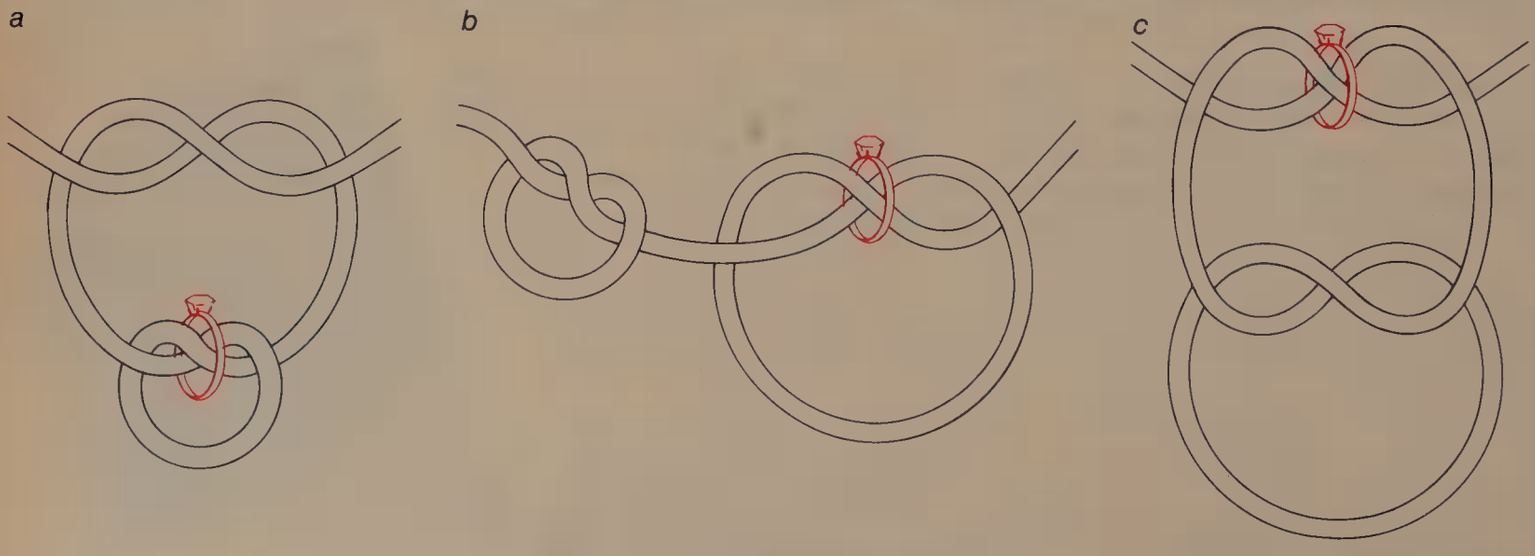
Can the Tower of Hanoi be solved by completely nonalgorithmic methods? Can it be done with a spreadsheet? Exhibiting such a solution would certainly not be a proof that a spreadsheet can do anything an algorithm can, but it would considerably enlarge the spreadsheet's range of action. Note that there is a trivial method of solution that must be declared out of bounds. One can solve the problem by hand, noting the configuration of the disks at each stage, and then write a series of formulas specifying the transitions from one configuration to the next. It is characteristic of such forced methods that with any slight change in the initial conditions, such as the addition of another disk, one must essentially start over. A robust solution would be able to handle any size stack without change or perhaps with a change only in the dimensions of an array. Note also that at least one spreadsheet program, Lotus 1–2–3, includes a simple algorithmic language; clearly all recourse to this facility must also be ruled out.

Another interesting case is the eight-queens problem, where the task is to place eight queens on a standard chessboard so that no queen is attacked by any other. Here the format of the problem—the finite array of squares—is tantalizing. There is certainly no difficulty representing a chessboard with a spreadsheet program. It is also easy to write a formula that reports whether a cell is currently attacked by a queen elsewhere on the board. (The formula merely checks for a nonzero value along all columns, rows and diagonals for a distance of eight cells.) If this were all one needed to solve the problem, however, it would not have attracted the attention of Carl Friedrich Gauss, who investigated it in 1850 but did not solve it. It seems each cell needs information not only on the current configuration of the board but also on the record of configurations that have already been tried. The difficulty of supplying this information in a static representation of the problem suggests that algorithms have a secure future.

The two knot problems posed last month by Martin Gardner are answered as follows:

The top illustration on this page shows how a square knot can be changed to an alternating knot of six crossings. Simply flip dotted arc *a* over to make arc *b*.

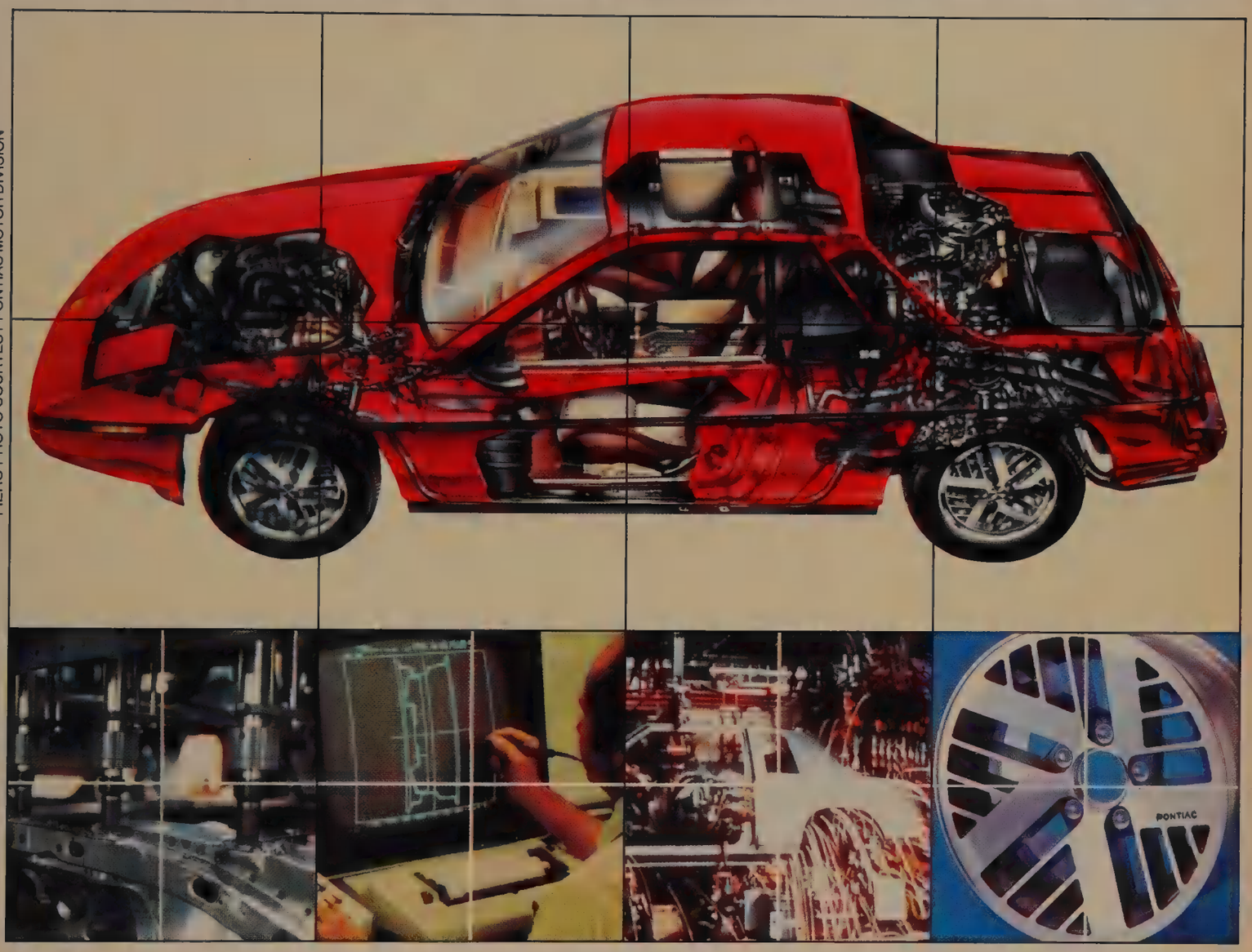
The illustration below shows one way to solve the ring-and-granny-knot puzzle. First make the lower knot small, then slide it (carrying the ring with it) up and through the higher knot (*a*). Open it. Two trefoil knots are now side by side (*b*). Make the ringless knot small, then slide it through and down the other knot. Open it up and you have finished (*c*).



Solution to the ring-and-granny-knot puzzle

PARTNERS IN QUALITY

DETROIT AND ITS AMERICAN SUPPLIERS



FIERO PHOTO COURTESY PONTIAC MOTOR DIVISION

This special report was prepared by
Jim Plegue of *Automotive News* in association with
Development Counsellors International, Ltd.

SCIENTIFIC AMERICAN

ADVANCED TECHNOLOGY CENTERS USA ADVERTISING SUPPLEMENT/SCIENTIFIC AMERICAN/OCTOBER 1983

INTRODUCING ONE OF THE MOST INNOVATIVE AMERICAN CARS EVER.

PONTIAC FIERO





Fiero can be described as an economy* car. It can also be considered a practical and durable car. And, by its technical specifications alone, Fiero can be legitimately called a sports car. The beauty of Fiero is that it's all of these cars. Best of all, Fiero is one of Pontiac's best examples of what driving is all about: pure unabridged fun!

Fiero is America's first two-seat, mid-engined production car. Fiero features fully-independent suspension, rack and pinion steering and power four-wheel disc brakes, all standard. Fiero is powered by a responsive 2.5 liter 4-cylinder engine with electronic fuel injection.

Fiero is designed to cope with the realities of driving in the real world: Its Enduraflex™ panels will never rust, the front fenders "give" on minor impacts, all panels are "ding" resistant, and Fiero's clearcoat paint wears a gloss that rivals the world's most pampered show cars.

Fiero! Not only a proud new American car unlike anything else in the world, but a dazzling display of Pontiac innovation in action!

*Pontiac Fiero Sport Coupe with available automatic transmission offers an EPA EST MPG of (27) and a highway estimate of **40**. Use estimated MPG for comparisons. Your mileage may differ depending on speed, distance, weather. Actual highway mileage lower. Some Pontiacs are equipped with engines produced by other GM divisions, subsidiaries or affiliated companies worldwide. See your Pontiac dealer for details.

GM

PONTIAC WE BUILD EXCITEMENT

Detroit is a mechanism for the manufacture of automobiles. In the early Seventies the mechanism worked smoothly in an environment of low energy costs, modest government regulation and relatively low labor costs. It made products which American car buyers believed of great value—greater than anything produced by a mechanism operating from abroad.

Then the environment changed, and the mechanism ran rough. Suddenly OPEC upped the cost of crude oil by what would become a factor of five. The United Auto Workers obtained greater wage-benefit increases just as sales and profits peaked. The Federal government added minimum Corporate Average Fuel Economy regulations to its safety and pollution-control standards. And Japan, Inc. introduced a line of low-cost, high mileage autos, products of a "mechanism" that had been revolutionizing itself even as Detroit's mechanism was showing signs of strain.

But Detroit—the auto manufacturers, suppliers, workers, and support infrastructure in finance, insurance, state government—is a rugged and resourceful super-industry, willing and able to do what it saw must be done. Detroit was determined to become America's leading advanced-technology industry. The new environment required it: cars had to be made to emission-control, safety, and fuel-efficiency standards largely set by outsiders. These changes would, in effect, "retool" the nation.

Detroit's network of suppliers, and suppliers-to-suppliers, reaches throughout the continent, with assembly plants in Kentucky and California, glass, rubber, and steel makers in Pennsylvania and New Hampshire, chemical resources in Connecticut and Delaware, all developing, manufacturing, transporting and installing some portion of a vehicle. New ideas from Michigan would, and must, require reindustrialization in hundreds of cities. The automakers planned and decided. The turnaround began almost at once.

IN SEARCH OF QUALITY

The original Detroit mechanism was based on the assembly-line: many cars were made at a minimum cost-per-car. Quality was external—not absent, but not integral to the mechanism. Thus, in the late Seventies the mechanism began producing vehicles similar to Japanese models, and at similar cost, but of inferior quality. Every calculation showed that increasing quality to acceptable levels would increase costs to unacceptable levels. The mechanism wasn't working, and no one was repairing it.

Then, in 1980, an NBC-TV documentary entitled "If Japan Can . . . Why Can't We?" was seen by William Hoglund, general manager of GM's Pontiac Motor Division, and some of his key associates. The program, and the processes it suggested, had substantial impact at Ford and Chrysler as well as GM. It displayed nothing which hadn't already been put forward regarding integration of productivity and quality. But its timing, and its audience, were just right.

Ironically, the Japanese mechanism was working well due to the work of an American, Dr. W. Edwards Deming, who had established a reputation in this country for his work in statistical sampling. He had been asked by a Japanese science group to address them on the subject of application of his methods to certain Japanese quality-control problems. According to Deming, achieving higher quality would improve productivity by reducing rework and defect-replacement. The Japanese adopted several Deming concepts, and Deming came very much to the notice of management in Detroit. Deming was hired as a consultant to Pontiac, to Ford, and to other industry producers and suppliers. As Pontiac general manager Hoglund said in *Fortune,* "His message shook the foundations of our approach to quality." John Manugian, executive director of product assurance at Ford/North America, met with Deming often, and Ford's production lines reflect many of the Deming methods.

THE DEMING METHOD: QUALITY TECHNOLOGY

At Pontiac, Deming detailed his 14-point special program for the virtual reconstruction of American car manufacture:

1) Create constancy of purpose toward improvement of product and service, with a plan to become competitive and to stay in business. Decide whom top management is responsible to.

2) Adopt the new philosophy. We can no longer live with commonly accepted levels of delays, mistakes, defective materials and workmanship.

3) Require statistical evidence that quality is built-in. Eliminate need for mass inspection.

4) End the practice of awarding business solely on the basis of price. Eliminate suppliers that cannot qualify with statistical evidence.

5) Find problems. Management must work continually on the system (design, incoming materials, machinery).

6) Institute modern on-the-job training.

7) Institute modern worker supervision. The responsibility of foremen must be *quality.* Improvement of quality will automatically improve productivity. Management must act on reports from foremen concerning barriers such as inherited defects or machinery not maintained.

8) Drive out fear, so that all may work effectively.

9) Break down barriers between departments. People in research, design, sales and production must work as a team.

10) Eliminate numerical goals and slogans for the work force, which demand new levels of productivity without providing new methods.

11) Eliminate numerical quotas.

12) Remove barriers between the hourly worker and his pride of workmanship.

13) Institute education and retraining.

14) Create a top management structure which will push every day on the first thirteen points.

For many in Detroit, the thought of implementing the Deming principles came as a shock. They were too much against the grain of "business as usual." Said Clayton Williams, manufacturing superintendent at Pontiac's Engine Plant No. 18, "At first, I didn't know what to think because the things he was saying like, 'Don't accept any defects,' were so radical compared with the way we had been running the plant. We lived with defects every day. To me it was something we couldn't do. But after being to seminars and sessions where he would speak, after the third or fourth time it occurred to me that what he was saying was attainable and could be done.

"I really believe and support those 14 points 100 percent. They work."

Several of Deming's points concern implementation of Statistical Process Controls (SPC). These work as follows: at each "process" of production, samples are measured for variation within each batch of parts, and then each batch is measured against design specifications. If samples are within predefined tolerances, production continues. If not, the production line stops, and the cause is determined.

Utilizing SPC to define acceptable

ADVERTISEMENT

quality at each individual process in the production path allows any problem to be rectified quickly. Ideally, it does so all the way through the supplier chain, because if SPC is to work, it must monitor the course from raw material to the finished part.

In the plant itself, SPC has yielded good results. Pontiac manufacturing superintendent Williams cited an example of a worker charged with boring holes in engine camshaft gears. Improper work—and tolerance here is just a few thousandths of a centimeter—means the gear will be loose on the camshaft; the engine will be destroyed in use.

Before the implementation of statistical process controls, explains Williams, "the employee would come to work, turn on his machine and check one part. If it was within spec, he'd turn the machine on and let it run for, say, two or three hours. If the part wasn't within print specification, he'd make an adjustment to the machine and run one more part. If that was okay he'd let the machine run for the rest of his shift."

Under SPC, the same worker would check three camshaft gears per hour, instead of one per shift. And if any one of those gears was outside tolerance levels, he was required to adjust the machine and check three more until the machine would properly produce three.

"SPC showed the employee that he would unknowingly have been making bad parts," says Williams, "because the first part was good, but the second and third were off spec. The three-piece check gives us a confidence level of, say, 95 to 99 percent that all parts measured within that hour are good."

In its first full year of implementation, SPC reduced the number of engines found defective due to cam-gear problems from 38 to six. Through the first half of the 1983 model year the number of failures was *zero*. Pontiac was succeeding: process by process, part by part, high quality was becoming not a goal but a reality.

The evidence mounted. Material Complaint Notices (MCNs)—complaints by Pontiac inspectors against suppliers shipping defective parts—showed an overall 15.8% defect rate for the 1982 model year. Through the first half of the '83 year, the identified rate was just 1.6%. Meanwhile, as defects dwindled, productivity skyrocketed—just as Deming had predicted. Says Pontiac's Williams, "Total productivity improvement averaged 25.6 percent in the 1982 model year. We expect another thirteen percent this year and another seven percent in 1984. If you add it, you're talking about a fifty percent improvement within a span of two to three years." A Ford Motor Company spokesman documented similar defect-rate improvement.

PARTNERSHIP FOR QUALITY

Improvements are attributable to more than SPC alone. Pontiac has implemented a series of new practices based on Deming's ideas, several aimed at improving the supplier/manufacturer relationship. Pontiac and other manufacturers have worked to create what Pontiac calls "an up-front partnership" with all of its key suppliers, based on a spirit of teamwork and cooperation.

Early sourcing, or "pre-sourcing", is one new practice. "Early sourcing" means a supplier is selected before a part has been committed to anything but a preliminary sketch, usually based on a computer-assisted design. High quality, not cost, is the primary criterion. Inputs from all affected staff areas help buyers make informed decisions. Early sourcing also allows the suppliers' R&D people to become involved in the design of a part to ensure that it is compatible with manufacturing processes.

Sourcing families of parts improves quality and streamlines contacts with suppliers. For example, headlamp doors and hood for Pontiac's new 2-seat mid-engine sports car were sourced with a single supplier because the "fit" of these parts is critical. It was thought that a single supplier could better control a group of close-fitting body panels for Fiero than a number of suppliers working independently. For the same reason, several body panels and covers, fenders, and doors were sourced together.

Single sourcing is just what it implies: selection of a single source to supply total production of a part. Historically, manufacturers selected several sources to keep costs low through competition, and through fear of over-dependence on one company. Deming allayed the latter concern with an analogy: two symphony orchestras play the same piece of music with equal fidelity to the notes on the page, but the performances differ because sheet music dictates only so much. Neither orchestra plays the music incorrectly, but results will be different all the same. An engineering "blueprint" is an automotive supplier's sheet music: give two suppliers identical blueprints and chances are that without either supplier deviating from specifications, the parts will be subtly different. As Pontiac director of materials management Donald Pais explained, "Once you develop a source who can meet your quality expectations, what would be the driving reason to go to a second source and duplicate the same effort on the same part?"

At the outset of the Fiero program, top Pontiac management met with top management from key suppliers under an *early contact* procedure. This let suppliers know in advance what would be expected of them. In these early meetings, Pontiac managers stressed the importance of quality in the finished car and its individual parts, and presented an overview of engineering design, manufacturing processes and expectations for service, quality, and delivery.

Pontiac then implemented "one-on-one" meetings between suppliers and buyers, including design and reliability engineers. Discussions of such concerns as identification of critical tolerances were aided by the greater specificity of the Deming methods. If close tolerances were unnecessary, every job became easier. Fred Seng, Pontiac's manager of supplier quality activities on the Fiero project, observed, "We found out—it amazed us—that we were actually asking our suppliers to make parts to certain tolerances that really weren't required. A hood that is, say, a mil and three-quarters wider than what the engineer originally designed really does make a good-looking car. We really don't have to sort that out, inspect for it and throw it away. The drawing says it's not quality because it's oversize, but the car says it's a quality part, and that's really the Bible. The car and its function and its looks really has to be the Bible." To date, nearly 100 "one-on-one" meetings have occurred.

A third stage of Pontiac's effort to improve the supplier-manufacturer partnership included the Fiero Supplier Seminar held in November 1982. 250 suppliers heard management and union officials stress quality and teamwork. At the Seminar, Fiero suppliers saw how their respective parts were assembled on a finished model. And in a developing fourth stage of its "partnership" program, Pontiac seeks to establish daily contact among purchasing agent, assembly plant, and supplier, including regular meetings and tours of facilities.

Comments materials management director Pais, "Typically the process had been to have a single part come in

from a supplier. It was laid out, given dimensional checks—and approval.

"We've changed, in that the product development team stays right with the source on a given part number, right through every design change. It's a different type of approval process, a very significant departure from the way we did business."

The new partnership between Detroit and its suppliers is wide-ranging. At Ford, for example, classes are regularly held for suppliers having difficulty with statistical process controls. "We try to bring them back up to speed," says Jim Olson, a manager in Ford's Detroit region. "We're reluctant to drop a supplier that falls below the mark until we've had a chance to work with him."

Elio P. Lori is plant operations manager for Ford's Engine Division. He cites a personal example: "Best efforts—trying hard—doesn't do it. We had to drop one supplier for consistent poor quality. He then installed SPC and convinced us he had relearned his company's process better than he ever knew it before, and we took him back."

Cost-savings have come about through adoption of Japanese "kanban" or *just in time* (JIT) inventory controls. This process streamlines production by reducing inventories from boxcar quantities to minimum quantities necessary to maintain assembly schedules. Donald Abelson, Pontiac's general superintendent-manufacturing engineering, describes the effect of JIT: "We took several major commodities—engines and sheet metal parts from our plants—and we have them delivered on a daily basis. Some suppliers like Lear Siegler (maker of Fiero seats) will be delivering three times a day. JIT translates into less 'bank' here at the plant, less inventory, which also says that you're going to turn over materials faster."

Traditionally, a supplier would manufacture and ship several months' worth of parts in advance. If a running change were made to the design of a part (as is common), the revised part could not be incorporated until after the existing supply of now-obsolete parts was exhausted. Today, daily shipments by suppliers means that a design change made on Tuesday might be in the car by the weekend.

COST OF QUALITY

Those at Pontiac involved with pricing have confidence that SPC and other Deming principles are working as planned. As Pontiac's Seng explains, "What you've really got to understand is that the cost of quality really is the expense of doing things wrong. If you consider all facets of part/supplier/product development under this team approach, you're going to save money because you're going to avoid making mistakes—and the mistakes are where your money goes."

Tom Charlton is primary purchasing agent for the Fiero program. About seventy percent of Fiero parts (on a cost basis) are being pre- and single-sourced, and it's Charlton's job to work with suppliers to achieve acceptable cost levels without undermining quality goals. Has his job become more difficult?

"Much more so," Charlton states. "In all areas. Because you're spending more time negotiating with suppliers over price and engineering changes. But we still feel it's the right way to do things. We've got to get away from price-buying, where you're always buying from the low bidder, because you're not necessarily getting your best value."

The actual cost per part, from the supplier to the manufacturer, may be higher under the new program than it would have been under older, cost-based methods. But such costs are offset by the superior quality of parts coming in and lower defect and scrap rates, not to mention Deming's all-important "less rework." Pontiac staff reliability engineer Donald Delaney predicts that after six months of operation, just the reduction in scrap rates may actually prompt cost decreases.

THE SUPPLY SIDE

Like a few other suppliers, U.S. Steel has been applying statistical process controls to its manufacturing operations for a few years. Bill Hughes, general manager of sheet products for the nation's largest steelmaker, says "We've definitely seen a change, a new openness between ourselves and Detroit. We are now given access to information that the auto companies considered much too confidential to release just a few years ago.

"That openness has allowed our own statistical process controls system, which was instituted in a very long, painstaking operation, to bring results. We're now enjoying some of the fruits of our investment."

Hughes points out that elements of the Deming method must be applied with care. "JIT is an example," he says. "If all it does is switch large quantities of material from a Detroit plant to a supplier warehouse, nothing's gained and the consumer pays. But an efficient flow of materials, accomplished through accurate forecasting and accurate scheduling, reduces costs. And that's when everybody truly wins."

Libbey-Owens-Ford is sole supplier of laminated and tempered glass for Fiero and is a major glass supplier for all of GM. Tom Walbridge, vice president of automotive OEM (original equipment manufacture), describes the L-O-F relationship: "On Fiero, we worked long in advance to help design the glass parts. In any car design you've got to work as much as four years ahead with the manufacturer's engineering staff, to ensure high optical quality.

"Statistical process controls aren't new," Walbridge adds. "We've all been using them more or less for a number of years. But SPC is being used more by *this* company, on *this* car, than it ever has been before."

For the future, Walbridge says, "We're getting into closer working relationships aimed at minimizing inventory on our customer's floor, and even more formalized programs of applied statistical process controls."

Dana Corporation utilizes an innovative approach to inventory-handling management. Don Decker, public relations director, explains: "We've built a network of small "satellite" assembly centers, all near vehicle manufacturing facilities. This is Dana's version of *kanban* . . . we're able to supply customer plants within 24 hours." Dana's own suppliers have begun using documentation of statistical process controls in lieu of inspections of incoming parts, and Dana provides such data to its customers—Caterpillar Tractor, for one—making receiving inspections unnecessary. Dana's people take pride in the recognition they've earned (including Ford's "Q1" supplier-quality award) as a result of their work.

"I like to think we were quality-conscious long before it became the thing to be," concludes Decker. "But we've put even more emphasis on it in the last couple of years."

Firestone's Tom Reese, vice president in charge of original equipment sales, says the firm is using statistical process controls in cooperation with auto manufacturers every day. Firestone is principal supplier of Fiero all-season radial tires, which it will deliver on a *kanban* schedule. "We're on the leading edge as far as just-in-time delivery goes," Reese notes, adding

that Firestone has maintained a 24-hour-a-day delivery schedule for two years at one Ford assembly plant in Kentucky.

Reese says the "up-front relationship" between Firestone and Detroit has had many significant benefits, one being an early decision to minimize tire sizes and types, allowing Firestone longer manufacturing runs, fewer changeovers and less down-time between runs. "The relationship is excellent, because it really does work both ways."

Fiero's is not the first non-metal body designed and produced in an American automobile, as anyone knows who recalls the 1953 Chevy Corvette. But most high-volume applications in Detroit tend to follow from earlier, more specialized applications. Now the Corvette may again be a precursor of a major production innovation: its 1984-model suspension and drive train, and many other parts, are engineered largely from aluminum and magnesium. The automobile is 200 pounds lighter, and thus more fuel-efficient, than its 1982 counterpart.

The Corvette represents manufacturer/supplier partnerships typical of the "new Detroit." After four years of R&D by Alcoa under Chevrolet's direction, strong, lightweight aluminum parts are fabricated by Martin Marietta, Dana, and several smaller corporations in Michigan, Ohio, Ontario. Supplier interest and cooperation thus merged with intensive new high-technology computer modeling and testing systems to allow a new process to begin. High performance and high quality have been attained with increased fuel-efficiency for the American consumer in the bargain.

The State of Michigan has been a major partner to the technological resurgence of Detroit's auto industry; the State, and its people, have benefited directly from Deming technology. Richard Cole, deputy director of marketing and public affairs at Michigan Department of Commerce, explains: "David Verway, the renowned automotive economist at Wayne State University here, predicts that just-in-time (JIT) inventory controls will increase supplier-related economic development activity throughout the State. Plus, by making training programs available and guaranteeing the excellent skilled work force which exists in Michigan, we were able to attract a major robotics equipment manufacturer to Sterling Heights, one which otherwise would have located elsewhere, perhaps overseas." Michigan has been particularly innovative in involving the State's educational system in the process of strengthening its economic base.

Martin D. Walker is executive vice president of Rockwell International Corporation, which adopted the Deming methods some time ago. "We use the term 'operator process controls' instead of 'statistical process controls,' " says Walker. "It may seem a small distinction, but in the final analysis the *operator* is the quality man, the guy who makes it go.

'The Deming methods surprised us. When we put them in, we quickly learned that our machines and processes were better than we thought they were. The more we taught our operators, the less we found ourselves adjusting our machines. We've tried to tell our own suppliers of the benefits of operator process controls. We tell them it doesn't cost — it *pays*.

"Operator process controls are important no matter what business you're in," Walker states. "We here at Rockwell, and the auto industry in general, believe in these methods. And you'll find that the automotive industry will soon be leading this country in terms of high-quality products."

Dave Denny is vice president/general manager of the EMS (transportation) Division of Premix, Inc., makers of Fiero's Enduraflex™ roof panel. Since their selection as a sole source, Premix engineers have worked closely with counterparts at Pontiac and Fisher Body Division to perfect the roof panel design. It is a process that Denny says has gone smoothly. "We've always wanted to have supplier input. I think it's terrific. I'm not sure I make any more money, but I prefer it. I like to know where I stand. I like to know my R&D dollars mean something."

Under the old system, Denny felt that sometimes quality suppliers were victimized by the low-bid process because they would spend money on research and design in order to build a quality part, then get undercut by a rival that spent less on quality. Denny also appreciates contracts for longer than the customary one-year term. "It's nice to know that, if I hold up my end, I'll still have the job several years down the road instead of having to bid on a year-to-year basis. It allows me to spend money and do the job right. I see a general trend in Detroit," Denny adds "I see a lot of emphasis on quality. And I believe it's real."

Jones & Laughlin Steel Corporation, a subsidiary of The LTV Corporation, manufactures many of the pieces that go into Fiero's all-steel space-frame. As at Premix, J&L engineers spent time with Pontiac engineers in materials development. Rich Cover, J&L's man-

The history of automotive technology as reflected in the pages of SCIENTIFIC AMERICAN:

A Gas-Propelled Carriage/Jan 1889

The Mass Production of Motorcars/Jan 1911

Balloon Tires And What They Mean/Jan 1924

The Fuel Problem/Dec 1949

High Compression/Feb 1950

The Electric Automobile/Oct 1966

Origin of the Automobile Engine/Mar 1967

Rotary Engines/Feb 1969

Rudolf Diesel and his Rational Engine/Aug 1969

The Great Automobile Race of 1895/May 1972

The Wankel Engine/Aug 1972

The Stirling Engine/Aug 1973

The Fuel Consumption of Automobiles/Jan 1975

The History of the Airflow Car/Aug 77

Alternative Automobile Engines/Jul 78

Reprints/copies may be purchased. Send check for $5/each to:
Manager—Industrial Sections
SCIENTIFIC AMERICAN
415 Madison Avenue
New York, N.Y. 10017

The science of quality in steelmaking.

U.S. Steel initiated its breakthrough Quality Management Program (QMP) in 1979 to assure optimum quality products for quality-conscious customers. QMP has already made a dramatic difference—with much more to come on a continuing basis.

At U.S. Steel, we recognized years ago that higher and consistent quality was literally a matter of survival—for ourselves and for our customers as well. It was this recognition that prompted us to establish our pioneering Quality Management Program.

Managing for quality.

This long-range commitment to make quality U.S. Steel's number one priority started at the very top level of corporate management and spread to include all levels. This continuing undertaking assures that production decisions of management are made primarily on the basis of quality, rather than tonnage. This is what we, at U.S. Steel, mean by "managing for quality." The logic is that "it pays to make it right the first time."

Technology for quality.

In order to assure that decisions for quality are sound workable business decisions, we began to educate our operating management in the new and exciting technologies for analyzing and controlling process variables—the "nuts and bolts" of the science of quality.

We are convinced that managing for quality and the technology for quality are necessary. Each is essential, but neither is sufficient in itself.

The technology for quality, as we see it, consists of seven key parts:

1. Develop in-process and finished product standards.

2. Identify critical major and minor process variables.

3. Control process variables to achieve desired levels of product conformance with reduced variability.

4. Develop systems to measure and report conformance to product standards.

5. Implement systems to measure and control processes to assure leadership quality.

6. Implement corrective actions.

7. Analyze performance and progress using cost of quality, product conformance and customer acceptance criteria as feedback.

Points 4 and 5 constitute the heart of the technology: statistical process control. But we believe that alone, they are not enough.

For example, producing quality products for the automotive industry requires advanced input in science and engineering. However, knowing what needs to be done and knowing how to do it are not sufficient. Consistent production of quality products requires an organized and continuing Quality Management Program approach. And ours is recognized by American industry as the most thorough and most successful in steelmaking.

QMP: how does it work?

Quality Management operates through working and steering teams. At USS Gary Works alone, we have 65 QMP working teams and eight division steering teams who select the working team personnel and priorities.

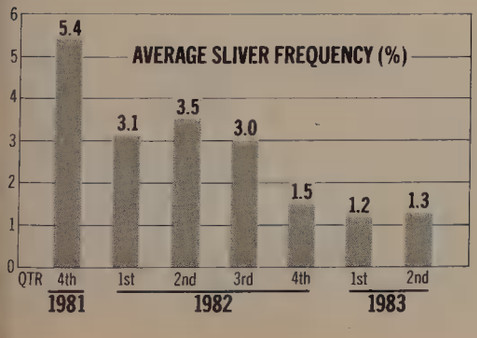
A typical working team consists of people from departments such as customer technical service, operating and maintenance, quality assurance, metallurgy, research, technical implementation, and, often, quality management consultants from outside U.S. Steel. Sometimes, the teams might also include marketing, accounting and industrial engineering personnel.

These multidisciplinary teams at Gary have played a major role in improving the surface quality and formability of steel sheet produced for the automotive industry. We have identified a number of critical product attributes which affect sheet steel performance. Included are chemical composition, surface texture and appearance and forming response. Here are but two examples of how our QMP at Gary has worked for our automotive customers.

Sliver imperfections decreased significantly.

Slivers are randomly occurring surface ruptures, discolored lines or streaks which are unacceptable on automotive body panels. During the fourth quarter of 1981, 5.4 percent of the sheet produced for exposed automotive applications contained sliver imperfections. A QMP working team was formed to attack this problem.

The team statistically analyzed steelmaking and continuous casting variables for possible correlation with sliver occurrence. In addition, our research laboratory personnel used scanning electron microscopy to examine a large number of samples so the sources of slivers could be identified. This work resulted in widespread process changes starting with the blast furnace and extending to steelmaking and continuous casting, hot rolling and cold finishing.



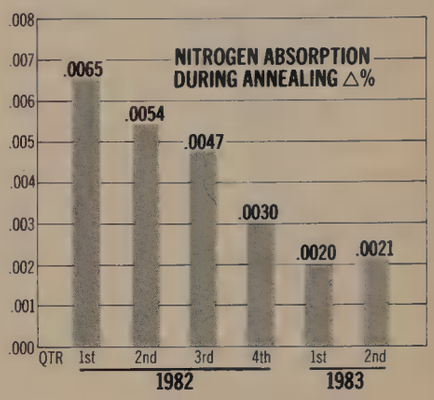
Reducing nitrogen absorption by two-thirds.

Early in 1982, a similar QMP working team was established to improve the formability of the automotive sheet produced at Gary Works. One of the goals was to reduce the amount of nitrogen absorption which occurs during annealing.

Cold rolled steel is annealed at elevated temperatures to develop the required level of ductility by recrystallizing the steel in a controlled manner. During annealing, nitrogen, which is a component of the protective atmosphere used to prevent oxidation, is absorbed by the steel sheet. Some nitrogen pickup can be tolerated, but at sufficiently high levels of absorption, free nitrogen will diffuse through the steel microstructure—leading to strain aging, which can adversely affect the formability of the steel sheet and the surface appearance of the formed part.

A significant challenge related to the task of the working team was to develop annealing practices which would reduce nitrogen absorption and still produce the desired recrystallized microstructure throughout steel sheet coils weighing an average of 25 tons. The work centered on two approaches.

One was to alter the protective atmosphere so there was less nitrogen absorption at the temperatures encountered during annealing. The other approach was to control the temperature gradients within the coils to minimize time and temperatures beyond those required.



These examples demonstrate how impressively effective our Quality Management Program can be. We are continuing to apply technology to further improve the quality of all our steel products. And we are relying heavily on statistical process control so that our customers can continue to benefit. We know, from several years of experience now, that assuring optimum quality is no "quick fix" or the simple adding of a few more hard-nosed inspectors. It is a full-time, all-out, long-range way of life. At U.S. Steel, it's here to stay and we think that's good for American industry.

United States Steel USS

TRADEMARK

ager of auto industry development, talked about the advantages he saw in early sourcing. "Before the auto companies did all materials development on their own. The supplier wouldn't be involved until the actual order. There wasn't any time for the supplier to volunteer his know-how. Any problem at that point was a major crisis. Solving it was a fire-fighting exercise.

"We think we can be a better supplier if we can be involved early in the design program. GM cannot be expert in everything. They need the expertise of the companies that serve them."

Cover cites examples of how working together helped to avoid problems and reduce costs on Fiero, and is encouraged by evidence that more manufacturers are moving toward "up-front partnerships" with suppliers. Like Premix, J&L is heavily into process controls and, according to Cover, already seeing positive results.

When asked about a possible dark side to Detroit's bright new ideas, one supplier paused for a moment, then replied almost apologetically, "I can't think of a single disadvantage."

Alan Scuterud is sales manager of The Homer T. Bronson Company of Beacon Falls, Connecticut. "Detroit's statistical process controls systems have challenged us to upgrade our own quality processing," he says. "We make hinges for Detroit, but we make them for plenty of other companies—Xerox, for example. What we've learned from Detroit has helped us serve other customers better."

THE SUPPLIER'S SUPPLIER

Mobay Chemical Corp. of Pittsburgh supplies many of the chemicals used in the manufacture of Fiero urethane body panels. Worldwide, Mobay's parent company supplies chemicals to Japanese auto makers, second only to U.S. auto makers in their use of the so-called "friendly" plastic. Having compared notes informally with some of his counterparts abroad, Dave Ahlgren, Mobay's automotive marketing manager, Polyurethane Division, reported that, surprisingly, the "team" spirit shown by U.S. auto makers may be greater than that of the Japanese.

"They don't appear to have the same cooperation we have here—that's an opinion, not necessarily a fact. We have an extremely good, close relationship with our customers here."

Mobay is fully committed to SPC and other Deming principles, reports Ahlgren, and is pleased with the results.

"We've always thought of ourselves as a good supplier, but we're learning to use the methods even more than we have, which, in turn, I think is making us an even better supplier.

"For those companies willing to make the commitment, it's a real plus. Those who don't probably won't be around."

Sheller Globe Corporation's Iowa City Division is Pontiac's source for Fiero's interior console—its dashboard components. According to senior vice president Sid Jeffe, the teamwork approach has been an unqualified success. "The difference is day and night as it relates to manufacturing feasibility," he says. "We have improved productivity in cycle time and manufacturing time."

Jeffe relates experiences from the past in which manufacturers would send design prints where the designer just didn't understand fully the properties of the material. Such mistakes would result in a costly redesign.

Jeffe reports great success with the concept of the use of "quality circles," which allows everyone in the plant to get involved with problem-solving.

Like other companies, Sheller Globe has had some difficulty convincing its own suppliers to implement statistical controls. But Jeffe believes this, too, will change. "All we're telling our suppliers (about implementing SPC) is that it's a matter of survival. It's that simple. It's one of the major ways this country is going to regain its position of preeminence in world industry."

Cover photography courtesy: Pontiac Motor Division; Rockwell Int'l Corp.; Michigan Dept. Economic Dev.

PONTIAC'S NEW FIERO OVERSEAS

Fiero may create, for American automotive products, export markets which have not existed for years. U.S. auto makers (according to the International Trade Commission) sell about 450,000 cars annually in foreign markets. About 350,000 are sales to Canada, which has no local auto industry and which maintains an auto trade agreement with the U.S. Middle Eastern countries account for much of the balance; Japan, which sells more cars to the U.S. than any other nation, imported about 3500 American-made cars—an insignificant amount indeed given that Japanese consumers registered more than three million new cars last year alone.

The U.S. automotive industry's export performance has been affected most by the "red tape" of special regulations. Stiff import duties and tariffs, plus local-content laws which require a high percentage of locally-made parts in imported cars, have hurt American exports. Then, too, American cars are designed primarily for American roads, tastes and needs. A large option-laden car simply doesn't function on narrow Japanese roads, nor does it offer the fuel economy necessary in nations where motor fuel can be twice the U.S. retail price.

Although virtually all General Motors cars today are designed with export possibilities in mind, Pontiac Motor Division's Fiero probably comes closest to addressing the problems which have plagued U.S. exports in the past. Because of its unique design, Fiero offers unique opportunities. Division officials have not disclosed their plans, other than to say there will be no Fiero exports during the first year of production. They do state a likelihood that Fiero eventually will be a sales factor in foreign markets.

Most important, the Fiero space-frame design allows far greater flexibility to meet foreign-market wants. Says Ron Rogers, director of the Fiero project, "Because of Fiero's construction, if you think about the body panels and the way they're put on the car ... you could do other versions or styles for an export market much more easily than you could with a car of standard construction." When assembled, the Fiero's space-frame is a rolling chassis, a completely functional automobile in all respects except that it has no exterior. This method of construction allows Pontiac flexibility to build a radically different car simply by designing new panels to fit the existing chassis. Slight changes could even be made to the space frame itself to accommodate a new "skin." A car—or cars—with different character can easily be created for export; at the same time, modifications are possible to meet a nation's special regulations.

It is conceivable that Pontiac could export only Fiero's rolling chassis to an overseas manufacturer, perhaps an affiliate or international partner of GM. The final automobile could be assembled in the receiving country, using Pontiac-made body panels or panels supplied by a local manufacturer. In this fashion Pontiac/GM might satisfy requirements currently leading to import duties or local-content regulations.

Fiero may indeed open up a "whole new world" of business for Detroit.

BOOKS

Ecological energy, round cities, the goddess Inanna, Atlantic fishermen, the robot family

by Philip Morrison

ENERGY AT THE SURFACE OF THE EARTH: AN INTRODUCTION TO THE ENERGETICS OF ECOSYSTEMS, by David H. Miller. Academic Press ($54.50). It does not require the cunning devices of double entry to be quite sure that the account books of energy are in neat balance. That budget is scrupulously managed by a higher law. This comprehensive and interesting text is a study of the energy budget of living systems in the form best suited to dynamic life spread out under the sun: the chief currency we count is in fact power density, watts per square meter of ecosystem area. (The heat supplied by an ordinary electric blanket flows at a power density of about 30 watts per square meter.) The ecologist, however, is no auditor: it is not that he is trying to balance the books the better to compute the bottom line. The abstraction of energy balance serves him rather to analyze the variegated grasslands of Denmark or the bed of a dry lake in California; the interest lies not in the sum but in its terms, some familiar, some surprising, together presenting one unifying aspect of the diversity of life on the earth. There is even a chapter mainly on the city, "a fascinating artifact; even if it were inhabited by ants rather than human beings, it would still be a worthy object of study."

The rays of the sun are the most important input. The solar beam is a changing vector, affected by night shadow and cloud, by the march of day and season, by dust and molecular scattering. Expect an average of a couple of hundred watts per square meter over day and night during summer in the Temperate zones, and below that by an order of magnitude and more in a European or Canadian winter. The beam outside the atmosphere rises to about 1,400 watts per square meter, beyond any reality on the earth. Temperature is a key consequence of the net flow of energies. One basic threshold is the freezing point of water, at which temperature these rough surfaces emit about 300 watts per square meter; at the higher-temperature end an extensive sunlit surface of water cannot get hotter than the low 90's (on the Fahrenheit scale), limited by its evaporation.

But it is realistically rich detail, not simple numbers, that gives substance to this set of some two dozen chapters. They are firmly based on worldwide field measurements of what really happens, expressed with plenty of tables and formulas; the text is not easy and is somewhat given to jargon of the trade, but it is without difficult mathematics. If sunlight has important changes in time and direction, it also varies in photon energy and hence in its effects on ecosystems generally, on the eye and the green leaf in particular. The solar infrared serves only to heat and perhaps elongate plants: 51 percent of the radiant flux is in that band. The ultraviolet is less than 10 percent. The rest we can see and may in part feed on. Animals bask in the direct sun; some arctic insects warm their wing muscles in the focus of the beam formed by the concave petal disk of the arctic poppy. But the sky diffuses the sunbeam, particularly under overcast or cloud; that hemisphere of sky light is complex, variable and important.

"In a way it is a source of wonder that the atmosphere, as deep and massive as it is, transmits any energy" down to the surface. That is of course the mark of the wide-spaced quantum levels of the major diatomic molecules of the air. In fact, there is a large and rather steady invisible flux of deep infrared we seeing creatures tend to overlook. It comes from the air above warmed by the copious solar infrared, and it amounts to as much as 400 watts per square meter, even more when water droplets high above join in. It is the rare trace molecules with three atoms that supply this radiation: most of it is contributed by water and carbon dioxide. In the wavelength range from four to eight micrometers the flux is local, coming from the lowest 50 meters or so of the air.

Most of this energy flux comes in at still longer wavelengths from parcels of trace gases the better part of a kilometer up, and a little comes all the way from the ozone high in the stratosphere. No plants feed on this too soft photon stream, but its role in their overall heat balance is clearly major. What doth it profit a leaf to nourish its chloroplasts if the entire structure is destroyed by

INVEST YOURSELF



A windmill to pump water for "salt farming" in India. More efficient woodburning stoves for the Sahel. Photovoltaic irrigation pumps for the Somali refugee camps

All these are solutions to technical problems in developing countries. Devising such solutions is no simple task. To apply the most advanced results of modern science to the problems of developing areas in a form that can be adopted by the people requires the skills of the best scientists, engineers, farmers, businessmen—people whose jobs may involve creating solid state systems or farming 1000 acres, but who can also design a solar still appropriate to Mauritania or an acacia-fueled methane digester for Nicaragua.

Such are the professionals who volunteer their spare time to Volunteers in Technical Assistance (VITA), a 20 year old private, non-profit organization dedicated to helping solve development problems for people world-wide.

Four thousand VITA Volunteers from 82 countries donate their expertise and time to respond to the over 2500 inquiries received annually. Volunteers also review technical documents, assist in writing VITA's publications and bulletins, serve on technical panels, and undertake short-term consultancies.

Past volunteer responses have resulted in new designs for solar hot water heaters and grain dryers, low-cost housing, the windmill shown above and many others. Join us in the challenge of developing even more innovative technologies for the future.

VITA **Putting Resources to Work for People**

1815 North Lynn Street, Arlington, Virginia 22209-2079, USA

CURE FOR THE COMMON COLD.



Bitter cold strikes young and old alike. And about all you could do before Thermolactyl was relieve the symptoms.

You could try to block the icy winds with layers of bulky clothing. Or wear old-fashioned long johns and perspire the moment you step indoors.

But now Damart is offering a cure you can live with (and afford!): Form-fitting underwear made of Thermolactyl, a manmade fiber much warmer than cotton or wool and also much lighter!

Damart Thermolactyl retains your body heat. And yet it b-r-e-a-t-h-e-s, allowing perspiration to evaporate. So you feel warm and dry, not cold and clammy. You're comfortable outdoors and indoors, too.

No wonder more than 1,000,000 Americans already wear Damart Thermolactyl for outdoor sports or to prevent nagging aches associated with cold and dampness. No wonder it's the choice of polar and Himalayan expeditions, the Ladies Professional Golf Association, and the Jets, Patriots, Bears, Bills, Steelers, Eagles and Packers football teams.

Damart's colorful catalog features form-fitting tops and bottoms in all sizes (including tall) for both men and women. You'll also find caps, glove liners, socks, even thermal lingerie!

Send for your catalog right away. Not only will you survive the common cold; you'll thrive in it!



Damart

Mail to: DAMART, Dept. 34095
1811 Woodbury Avenue
Portsmouth, N.H. 03805

FREE CATALOG Yes—rush your catalog of warm but lightweight Damart Thermolactyl. I understand there's no obligation at all.

Name _____
(please print)
Address _____
City _____ **State** _____ **Zip** _____

WARMTH WITHOUT WEIGHT OR WETNESS.

A LIMITED WARRANTY TO CONSUMERS ★ Good Housekeeping PROMISES REPLACEMENT OR REFUND IF DEFECTIVE

© 1983 Damart

freezing? This long-wave radiant income is always compensated for in large part by the upgoing thermal infrared emissions from the ecosystem at the surface temperature.

About a third of the book deals with this subtle radiation problem, ending in an overall balance struck among seven different fluxes: the solar direct beam and the sky income at short wavelengths, the infrared from gas and cloud, and the partial reflections and emission upward. Capable all-wave radiation sensors in national networks have been in service for years in many countries, with 100 or more deployed across the U.S.S.R. One feature of this text is its excellent coverage of the literature, with much space given to the extensive Russian and Western European work.

The second third of the text goes beyond the radiative base. It treats of the carbon fixation that sunlight spins out of carbon dioxide and water: the energetic harvest. In the end that store is broken down by grazers, decomposers and fire. From leaf-cutting ants to the human harvester every visible feeder is unimportant quantitatively. Even the energy fixed in a field of wheat goes to sustain the unseen and the lowly; they take six times the amount cropped as human food. Visible grazers become energetically important only rarely, as in times of a caterpillar population explosion; then the litter fall from tree leaves can approach the usual values of net primary production, about one watt per square meter.

A merino ewe is diagrammed in her energy environment in the sunbaked Australian shrubland; her thick fleece physically rejects most of the solar input by long-wave emission. The shorn lambs endure an opposite stress; they depend on a tempered wind. In winter their heat loss can easily become lethal. Now the subject matter gets more diverse. There is latent heat from the phase changes of water and sensible heat to worry about, in transfer to the air and the soil. There is the heat content and even the kinetic energy of hail, wind and rain, often important as trigger processes. There is stratification: the tree canopy sees radiant flux and turbulent air quite unlike those of the forest floor. Not much is ignored; many of these items are small in the mean but serve to turn natural valves, so to speak, for the larger flows.

Finally enter the inputs from our species, by fire and tractor, by fertilizer and reaper. The closing chapters seek integration, and some careful accounts have been given. The text ends with a remarkable diagram, plotting on a triangular field the three major nonradiative fluxes, that of latent heat, sensible heat (given to the air) and soil heat, expressed as fractions of the net all-wave radiation, the three corners all idealized pure cas-

es. Desert ecosystems plot along the edge between the soil-heat and sensible-heat corners; the farm plots are watery enough to dwell near the latent-heat corner. There is an entire world implied: cactus stems, deer, bogs and the typical citizen of Christchurch, New Zealand, have all been modeled in such an energy-flow domain. Scale is important; the energy-budget method has been widely applied to food and feed crops, from an initial effort at single leaves. The hotlands sheep is a radiator, but insects are convectors, "keeping cool purely by virtue of their size."

The city treatment is brief but provocative. One result worth citing is the effect of the coming of electric power about a century ago. The cities' structure changed; they got much less intense (compare the inner heart of a walled city with the present sprawl), but the energy conversion needed for circulation has compensated for the increased area, and the overall flux density has remained at about 20 watts per square meter. It is not only fossil-fuel use that contributes: the city is an investment in energy-rich matter. Amortized over 40 years, the material inventory requires an energy flow about as great as fuel use. The heat island over cities is well established; its causes and changes are not yet fully known. Cities warm the night because they store heat underground; they give off more long-wave radiation, and they disturb the atmosphere both compositionally and thermally. The manifestly powerful world of furnaces, cooling towers and smokestacks lies in quite another regime.

CITIES IN THE ROUND, by Norman J. Johnston. University of Washington Press ($29.95). Eye-catching illustrations ornament and vivify this delightful study of a struggle between the ideal and the real nearly as far-flung over time and space as settled humanity itself. First a word of explanation for the title is needed. The circle is a precise concept, but its infinite symmetry has been seen ever since the great Florentine Leone Battista Alberti as only the limit of that considerable group of finite rotations in the plane. "It is manifest," he wrote, "that Nature delights principally in round Figures.... We find too that Nature is sometimes delighted with Figures of six Sides." He went on to admit other polygons as well. So does our author, a professor of architecture and urban planning and design at the University of Washington. Symmetrical forms as intricate as snowflakes in all their reentrant concentricity dominate the visual impression of this handsome book; true circles are scarcer, usually implicit.

Put aside the recent pivot irrigation that grows clusters of big green circles across the flatlands of some modern farmers. Otherwise the translational de-

THE LADY DAVIS FELLOWSHIP TRUST

P.O. Box 1255 Jerusalem 91904, ISRAEL

Fellowships for 1984-85 at the Hebrew University of Jerusalem or the Technion-Israel Institute of Technology, Haifa

GRADUATE AND POST-DOCTORAL FELLOWSHIPS

ELIGIBILITY: Lady Davis Fellows are selected on the basis of demonstrated excellence in their studies, promise of distinction in their chosen fields of specialization and qualities of mind, intellect and character.
APPLICATION INFORMATION: Graduate candidates may apply during their senior undergraduate year or after they have undertaken study in a graduate school. Post-doctoral candidates may apply not later than 3 years after completion of their doctoral dissertation.
The Fellowships are tenable for one year but may be granted for another year.
The grant covers travel, tuition fees and reasonable living expenses.

VISITING PROFESSORSHIPS

are intended for candidates with the rank of Full or Associate Professor at their own institution. They are tenable from one trimester (or semester) to a full academic year.
DEADLINE: Completed forms in all categories must reach Jerusalem not later than December 1, 1983. Late applications will not be considered.
Candidates, including Israelis abroad, may obtain application forms at the above address. Please indicate category of Fellowship required.

To preserve your copies of SCIENTIFIC AMERICAN

A choice of handsome and durable library files—or binders—for your copies of SCIENTIFIC AMERICAN. Both styles bound in dark green library fabric stamped in gold leaf.

Files **Each file holds 12 issues. Price per file $5.95; three for $17.00; six for $30.00, postpaid.** *(Add $2.50 each outside U.S.A.)*

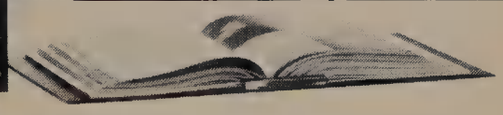
Binders **Each binder holds 12 issues. Issues open flat. Price per binder $7.50; three for $21.75; six for $42.00, postpaid.**
(Add $2.50 each outside U.S.A.)

To: Jesse Jones Box Corp., P.O. Box 5120, Philadelphia, Pa. 19141
Send me_____SCIENTIFIC AMERICAN ☐ Files ☐ Binders
For issues dated ☐ through 1982 ☐ 1983 or later.
I enclose my check or money order for $_____(U. S. funds only).



file binder

Name _____ *(please print)*
Address _____
City _____
State _____ **Zip** _____



NOTE: Satisfaction guaranteed or money refunded. Allow four to six weeks for delivery.

mands of the plow, like those of the carpenter and the mason, strongly favor the rectangle, or at least less regular ways of filling the plane. It is the circle as symbol that the round city celebrates, not any rotational way of work. It is true that the royal Zulu cities of the 19th century took the simple kraal to great dimensions: the king at the center, the warriors' huts ranged around him in four- or fivefold ranks to the scale of an imposing circular town of thousands. The Zulus were proud of a livelihood of cattle raising; their range was effectively unbounded, not divided by land deeds or barbed wire.

Other camp and village circles can be found, as among the nomad Cheyenne, or the gardening groups of the Amazon. Those peoples have neither plowed fields nor stone houses. Great China, and the Asian cultural domain it may have influenced or reflected, from the four-square pueblo to the stupa and the Purple Forbidden City, have a distinct cosmological need. Theirs is an earth of four cardinal directions; it is the sky alone that forms a celestial circumpolar circle. These shapes may be combined, as in a mandala or for the imperial ritual instruments of jade, but it has never been an Eastern plan to occupy this earth circularly.

Herodotus tells the tale of Ecbatana of the Medes, with its seven concentric walls in color, one for each luminary, from the golden central sun wall of the king to the outermost white wall for Jupiter. This is more likely to preserve myth than matter, but in fact there were circular cities in the ancient Middle East. One drawn here in aerial perspective is Gur of the Sassanians, built in A.D. 226 and left in ruin when Islam came to southern Iran. Its cosmic plan, centered on kingship, recalling the sun itself, persisted. Baghdad was planned by the caliph al-Mansur as a City of Peace and a paradise on earth; no form other than a circle would do, centered on the Golden Gate Palace of the caliph with its green dome. The fabulous city was built just so in four years. The purity soon departed, and development rapidly breached the ring wall. No sign of circles is found in Baghdad today.

The Classical world built no round cities; medieval towns were also walled orthogonally, although the City of God remained a circular ideal, theoretical and latent, "like embers." A remarkable example from that same time, but far from the culture of Europe, is the near-circular town of Mexcaltitán, on the west coast of central Mexico, built on an island in a lagoon. Its citizens hold that its form has cosmic meaning; they have retained the rounded wall and the central quartering by crossed streets with "singular resistance."

Theory caught up the Renaissance planners; theirs was a time of human reason. Domes burgeoned, and with the lofty circles of the central form came the idea of the perfect city. By the end of the 16th century, however, a strange alliance arose between philosopher and military engineer: the cosmic idea was linked to "a melange of trigonometry, cannon fire traceries and bombardment analyses." The prince at the center needed those convoluted bastions and lunettes; wide roads axially outward to the gates in the frowning multiple reentrant walls seemed necessary to quick reinforcement. The fortress town of Palmanova, the finest of the day, was sponsored by no single prince but by the rich Republic of Venice. Its ideal was fortification, and the swiftly realized plan, still more like a snowflake than any other city still standing, had to bend toward the practicality of a working city.

What any lesser prince could do, Louis XIV and Peter the Great would do even more grandly. The planned baroque town-forts multiplied from France to Finland; the great Vauban remodeled more than 160 fortresses and built 30 new ones, while directing 53 sieges. He was created Marshal of France, but his concern for the masses of the people and their taxation did not please the court; he died in disgrace.

The autocrats diminished, and artillery range outgrew the most ambitious walls. The new theories nonetheless retained a place for the city in the round, no longer to be justified by the centrality of a ruler or by the cosmic view. Now they were arguably pragmatic and egalitarian. (Theorists are up to a certain amount of rationalization in the interests of simplicity of form.) There are plenty of projects for such cities and a few realized ones, with more than one circle that was squared once the populations got their hands on the decisions.

A kibbutz called Nahalal is an open sunburst town, its fields radial strips out from the final ring of houses. Brasilia is no round city, but the runner-up for its planning competition of 1957 was a design by the brothers Roberto for a cluster of seven hexagonal urban units, housing 72,000 souls each, a plan that could expand by replication. Even later the workers who built the first stage of the power-station project on the river Paraná were housed in a circular town around a model farm. It was hoped the small city would take root. "But once construction ended, the residents hastened elsewhere." Poor Jupiá is a neat page in the architectural treatises, but the army took over half of the rings and the rest of the settlement was bulldozed.

Professor Johnston concludes that the persistence of the circle-city idea is a signal we can still make out amidst the noise of the shifting arguments for it. Somehow it is built into the mind and the heart. We need cities we can love, all right, but it is not the planners, however indispensable, whose harmonious emotional response to city form should stand first in priority.

INANNA, QUEEN OF HEAVEN AND EARTH: HER STORIES AND HYMNS FROM SUMER, by Diane Wolkstein and Samuel Noah Kramer. Art compiled by Elizabeth Williams-Forte. Harper & Row, Publishers ($16.95). At the eastern edge of the alluvial plain of the lower Euphrates in southern Iraq the archaeologists have for two generations been working at the mounds of Uruk, biblical Erech and modern Warka, one city-state of ancient Sumer. The site has disclosed marvels, the oldest written documents and the first great monumental temples, their mud-brick walls, huge columns and high platforms encrusted with a colorful patterned mosaic of cones and beakers of painted terra-cotta. The time of Uruk's flowering was a period of 500 years and more from about 3500 B.C.; Sumer endured, waxing and waning in the flux of human affairs, for another couple of millenniums.

The glory that was old Uruk was not forgotten: its beliefs and forms were recorded with learning and passion on clay. That very technology was Uruk's accomplishment. Here and there in newer cities throughout the land between the two rivers there accrued libraries of the hand-size tablets, marked in swift cuneiform by scribes and scholars who knew the past, even though they were of a different people. In the 19th century tens of thousands of tablets were found and preserved; it has been a triumph of our century to have translated many of them. Among all those temple accounts and tribute lists and school lessons we hold as well a literary treasure, "more than thirty thousand lines of text, mostly in poetic form, [constituting] the oldest written literature of significant quantity and variety so far uncovered." It includes "twenty myths, nine epic tales," 200 hymns, with laments, dirges and diverse texts of antique wisdom. The principal source tablets for the Inanna cycle are dated at about 1800 B.C.; their story is surely much older.

The poets and the people of Uruk listened and sang; the scholars of later Sumer recorded their liturgy in all piety; the dusty tells held the bone-old records safe until the artful diggers came; the scholars unraveled the crowded wedge signs into English, French and German texts. The lapidaries represented the sacred date palm and the Holy Fly of the Underworld, the nude Lilith and the radiant Inanna.

This volume of collaboration takes us one inviting step further. The young senior author is a gifted storyteller, an experienced folklorist, with a sense rare in our literate times of the demands and the rewards of the spoken tale. Searching for a grand story of a woman, she

Imported by The Paddington Corporation, New York, N.Y. • 34 Proof © 1983



FOR MAJOR CELEBRATIONS
AND MINOR VICTORIES.

BAILEYS® TASTE THE MAGIC.

came to realize that the tale cycles of Diana, Demeter, Cybele and Persephone and the others are all scattered and incomplete. It is Inanna, the oldest story of all we know, who alone spans the whole of the mythic life.

A girl, Inanna claimed her womanhood with the aid of her hero-brother Gilgamesh; she brought the sacred *me,* the root of all order, to her city Uruk to enter first on her queenship. She made joyous love with her chosen shepherd-king Dumuzi. Reigning, she journeyed to suffer the tyranny of the Underworld in death; she returned to the Great Above after exacting a partial sacrifice from Dumuzi. Finally she ruled Sumer in all her aspects, celestial and temporal, "mighty, majestic, radiant and ever youthful," patroness alike for sweet fruitful love and the bitter emptiness of death. Like Venus, she is together the Morning Star and the Evening Star, 2,000 or 3,000 years before the astronomers of Hellas.

The distinguished junior author, Samuel Noah Kramer, is the recognized dean of Sumerologists. His own scholarship has for 40 years given us most of the meaning we can now find in the tablets. He prepared the relevant Inanna texts in up-to-date form, some drawn from very recent translations and newly pieced fragments, taking account of all current work. Those texts, the best we have, he turned over to Diane Wolkstein, the storyteller. She has spent three years preparing her cycle of Inanna, more than 100 pages true to the resonant but unrhymed verse line of Sumerian. Not one tired archaism is left, no days of yore. She has shaped order and flow, keeping "as close as possible to the power, wonder and mystery" of the old texts. She has eliminated enough repetition to allow the story to breathe, and yet she has retained enough of it to give the novelty of a ritual authenticity long hidden.

Half of the book is this newly alive story of Inanna, half is a readable, varied and documented commentary from three points of view: that of the epigrapher, that of the writer and storyteller and that of the art historian and archaeologist. The gloss explains the complex written and visual material of this unusual and exciting volume, at a little-visited frontier between meticulous scholarship and vivid theater. Within the text there may well be lightly encoded one important constituent of the astronomical understanding of the time, only hinted at in this book.

"As Inanna ascended from the underworld,

The *galla,* the demons of the underworld, clung to her side.

The *galla* were demons who know no food, who know no drink,

Who eat no offerings, who drink no libations,

Who accept no gifts.

They enjoy no lovemaking.

They have no sweet children to kiss.

They tear the wife from the husband's arms,

They tear the child from the father's knees,

They steal the bride from her marriage home."

The hymns celebrate the bride's return in peace:

"My Lady looks in sweet wonder from heaven.

She looks in sweet wonder on all the lands

And on the people of Sumer as numerous as sheep....

I sing your praises, holy Inanna.

The Lady of the Morning is radiant on the horizon."

DISTANT WATER: THE FATE OF THE NORTH ATLANTIC FISHERMAN, by William W. Warner. Little, Brown and Company ($17.95). The drama is at the cod end. Studded with the heads of live fish bulging through the mesh, clouded over by screeching flocks of excited kittiwakes, it broaches heavily up in the wake of the trawler, also catching every fisherman's eye in that moment of visible triumph. The cod end is the socklike bag into which the caught fish collect as the big net is dragged through the sea. A trawler out of Boston on Georges Bank may net as much as a ton of fish in one lucky cod end; her company of 10 men reckon to spend a backbreaking week or so at sea between times ashore, and they will divide with the owner the proceeds of a catch of some 40 or 50 tons of well-iced cod and blueback pollock.

In competition with the bands of New England fishermen there arose a hunting system of a very different kind: an engineered enterprise of urban proportions afloat, taking fish at an order of magnitude greater rate, with 10 tons to the cod end, trawlers of 10 times the displacement and every crew 100 men—and women too. They spend months at sea, rotating by air to a home port as far away as Murmansk. Fleets of 200 such ships like cities afloat until recently drew their "big Jesus nets" across the Banks. This beautifully written book is an ethnography of that system, a loving, firsthand "foredeck chronicle" by a writer who spent months on shipboard among the enduring and skillful people of half a dozen nations who follow the last great hunting way of life.

The new way began with the *Fairtry,* a minesweeper converted in Aberdeen to fish the Grand Banks first in 1954. That British prototype launched 1,000 factory trawlers. (If we are to credit the account here, the first such Russian ships were built from her blueprints, cunningly wheedled out of her designers on a pretext of orders.) Four innovations defined the class: a stern ramp for the big nets, an automated fillet-making line sheltered belowdecks, a quick multiplate freezing plant aboard and fishmeal equipment to convert the part of the catch that is unmarketable as human food. Such a ship is big enough to offer movies and showers to the crew, and costly enough to justify powerful electronics, for example a set of sonars mounted on the elaborate nets, whose mouths may be 10 acres in area, guiding the mid-water trawl in three dimensions to envelop with precision the big sonar-detected schools. The quick-frozen morsels of white fish are nothing like the sallow week-old cod at the bottom of a dirty pile of ice that earn a longshoreman's grimace on the Boston Fish Pier.

They that do business in great waters earn our respect simply by completing their voyages; much more must be offered to those who also work hard and dangerously almost every hour at sea. We meet them through Warner's openly friendly sharing afloat. The West Germans are held to be the most expert. Their big instrumented mid-water catches do not stir the bottom but come up "clean"; regularly the cod end yields 98 or 99 percent of the fish, say herring, that they had trawled for. The fish-processing machinery is itself a worldwide near-monopoly of the Lübeck firm of Baader. History is plain on the fishing grounds; we ought not to be surprised that the best fishing in all waters and weathers is done from ships built for Hanseatische Hochseefischerei, nor that crews from new fishing nations, such as Cuba, show they have much to learn in distant cold waters.

It is on board a big Kaliningrad trawler that we glimpse something of the life of the Russian working people. Their trawlers are generously staffed, perhaps by half again as many crew as serve a larger West German ship; the frenzy and crushing fatigue of the work are gone, shared among many hands on a kinder watch system than the 12 hours on and six off of most fleets. Their captains cooperate, calling their colleagues to share a find instead of voicing the covert success and open deceit heard over the radio from the fiercely competitive masters of the West. The crews, men and women, are more diverse in career, ambitions and education than the tough fishermen of Grimsby or Boston; they go to sea for wages three or four times what a factory worker might earn.

We witness calisthenics and film after film of the Great Patriotic War, meet an interesting crew that serve unbonded by shared experience to any single ship or master and dine with them heartily, if predictably, on rich tasty soup to weak fruit compote in a mess hall decorated with eight hortatory posters. One reads in translation: "BALLS OF FIRE! The Brigade of Second Fishmaster A. T. Abybikhanov...has processed 5 TONS....

Helping engineers change gears.

The relentless ingenuity of America's engineers has made technological revolution commonplace.

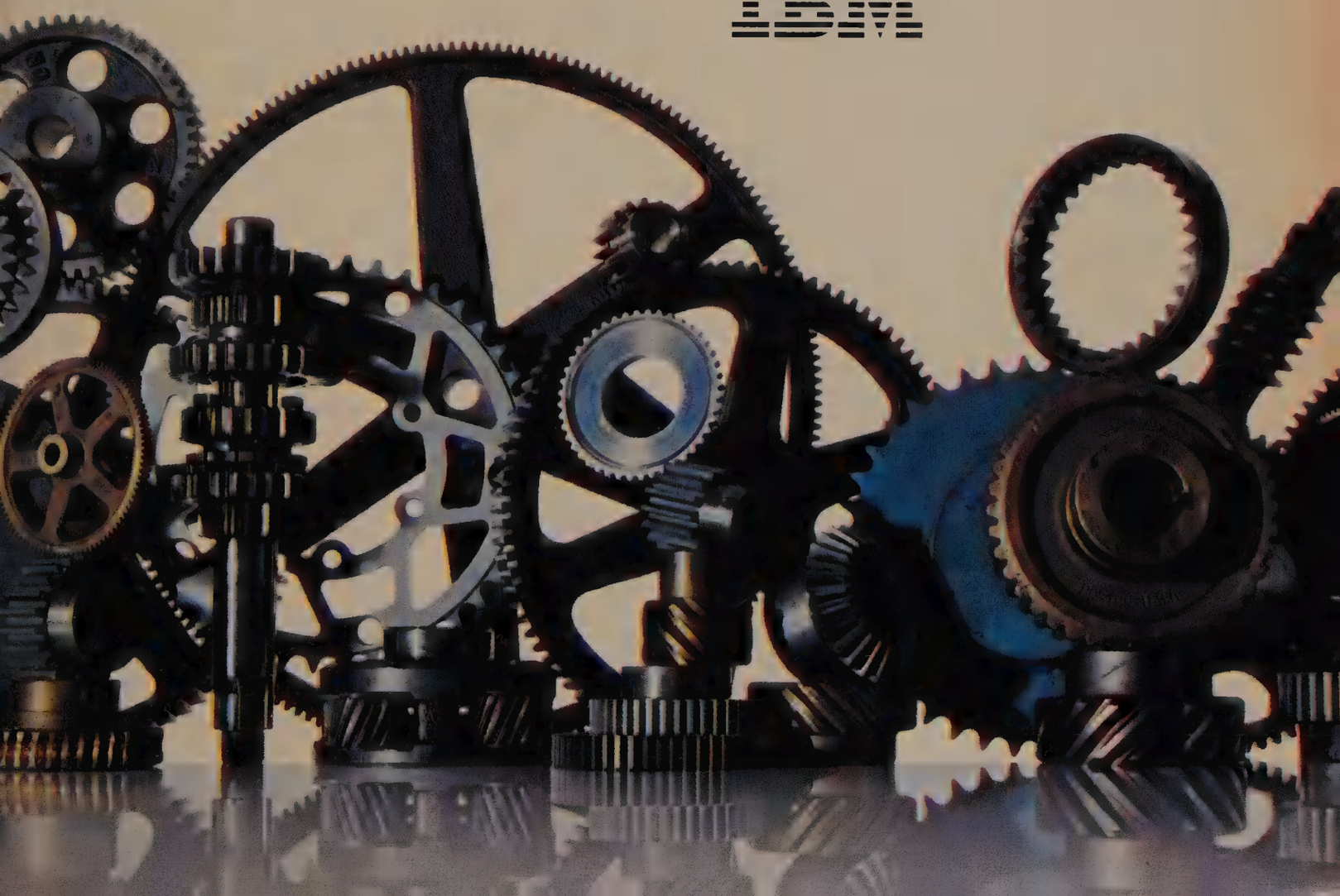
But today, manufacturing engineers face a relentless challenge: to produce goods of ever-higher quality at ever-lower cost. The marketplace, domestic and foreign, demands it.

To meet that challenge, engineers need state-of-the-art training and state-of-the-art tools. This, of course, is a job best done by America's universities. But universities can't do it alone. All sectors of society must help.

That's why IBM is giving $50 million in money and equipment to 22 universities for the creation and expansion of engineering programs in manufacturing systems. The curricula, in part, will include computer-aided design and manufacturing, robotics and lasers, quality control, systems management, and automation.

We expect this investment in education to pay off by helping to keep America's manufacturing in high gear.

And that will pay off for all of us. IBM®



BIG BOOM FOR REAL ESTATE INVESTORS.

Thousands of people like you find new opportunities for income, appreciation and tax advantages through the CENTURY 21® network.

Tax laws have changed. And now real estate is an even better investment than ever before.

At CENTURY 21 offices all over the country, we're showing people how to take advantage of the latest tax laws. How they can use leverage. How they can now shelter more of their income. Even income from salary, stock dividends and other sources.

A CENTURY 21 Investment Professional can show you, too. Through real estate, you may be able to get a greater annual cash flow, after taxes, than through other traditional types of investment. And you don't have to invest a lot of time and effort. Because a CENTURY 21 office can also handle the details of property management for you.

The following chart illustrates the potential advantages of real estate over other investments:

	REAL ESTATE	MONEY MARKETS	STOCKS
Cash Flow	✓	✓	✓
Depreciation	✓		
Loan Amortization	✓		
Appreciation	✓		✓

The opportunities are impressive. And that's one of the reasons investment business is booming at CENTURY 21 offices. Last year, we helped investors buy or sell real estate worth over $4 billion. We're North America's leading real estate sales organization. And now, we're leading the way in real estate investment, too.

Century 21®

Call this toll-free number for a CENTURY 21 Investment Professional.
1-800-228-3399

Nebraska Residents please call 1-800-642-8788. In Alaska and Hawaii call 1-800-862-1100.

©1983 Century 21 Real Estate Corporation, as trustee for the NAF ® and TM-trademarks of Century 21 Real Estate Corporation Printed in U.S.A. Equal Housing Opportunity

EACH OFFICE IS INDEPENDENTLY OWNED AND OPERATED.

STEADY AS SHE GOES!" Only Russian deckhands remain universal fishermen, complaining, bantering, horny and cynical. It was a happy ship, a safe one, its many jobs secure, and it evoked the air of a serious schoolroom, with grades and messages from the principal on constant display. The farewells to Warner, from the laundrywoman's gift of perfume to the captain's three hugs, left him with that long-familiar but still moving reflection: "To work and to live closely with the Russian people is to know generosity, warmth and a sense of community matched by none."

The factory fleets have departed now from the North Atlantic. The last of the big "floating cities" glowed offshore there some five years ago. In 1976 the U.S. and Canada led in the establishment of the 200-mile treaty limit on foreign fishing. Controls soon spread to all coastal states. The nutrient waters at the boundary of the Gulf Stream, just north of whose mixing flow the riches of the North Atlantic fishing grounds arch from Nantucket Shoals to Skolpen Bank eastward of the North Cape, support plenty of fish but not enough to sustain those megaton catches of the 1970's.

Nobody launches such big trawlers today; they are not well suited to the multiple species of the tropical Pacific waters whither the new fleets have moved. The British trawlers are laid up or are plowing Cornish waters. The Germans tried New Zealand, but it was hard to match the huge uniform schools of the cold North Atlantic shallows and the quotas fell swiftly. The Spanish pair boats that share one net drawn between them remain a small-scale exception. The Russian fleet has dispersed to the seven seas, often not fishing but merely processing at sea the erratic catches of local fishermen along the Benguela Current off West Africa, or creaking and groaning while they wait out sellers in the slow California swells.

The fish-bearing shelves of the world ocean are now nationally subdivided, "a staked plain"; the world catch remains about the same overall, but it is now far more diffuse. The North Atlantic shoals are livelier, not with big trawler fleets but with fish and birds. Only in the Norwegian Arctic do the catches still decline; elsewhere they are becoming stable, at levels higher than those of the past for local boats but severalfold below the transient peak of those too-well-supported hunters. The affectionate *nuk nuk* girls of Godthaab, like the Stardust Club of quiet St. John's, must miss their foreign fishermen nowadays. The Boston boats still arrive one or two at a time at the Fish Pier almost every day, each laden with 40 tons.

INDUSTRIAL ROBOTICS HANDBOOK, by V. Daniel Hunt. Industrial Press Inc., 200 Madison Avenue, New York, N.Y.



Magic Baloney

You'll love the way we hated the Magic Stat thermostat until an amazing thing happened.

It had no digital readout, an ugly case and a stupid name. It almost made us sick.

You're probably expecting our typical sales pitch, but get ready for a shock. For instead of trying to tell you what a great product the Magic Stat thermostat is, we're going to tear it apart. Unmercifully.

When we first saw the Magic Stat, we took one look at the name and went "Yuck." We took one look at the plastic case and said "How cheap looking." And when we looked for the digital readout, it had none. So before the salesman even showed us how it worked, we were totally turned off.

REAL LOSER

So there it was—at first blush a real loser. But wait, we did find one good feature—a feature that led us to a discovery. The Magic Stat installs in a few minutes and no serviceman is required. Thermostat wires in your wall follow standard color codes. So when you install Magic Stat, you attach the red wire to the red location and the white to the white. That's playschool stuff. And it's safe. Conventional thermostats installed over the past 20 years are generally only 24 volts, so you can either turn off the power or work with the "live" wires without fear.

OK, LET'S TEST IT

The Magic Stat installation was so easy that the least we could do was test it. And that's when we made an incredible discovery. We discovered that the Magic Stat was probably the most consumer-oriented, technologically-advanced and most sophisticated thermostat ever developed on the face of this earth and in our galaxy for all times ever. What made us switch from hating the thing to loving it? Read the following.

The Magic Stat has six setback settings per day and a seven day program. That means that you could set it for 70° when you get up in the morning, drop the temperature to 54° when you go to work, raise it to 68° when you return for dinner, raise it up to 72° after dinner as you watch TV and then drop it down to 62° when you go to sleep. Count them—five settings with one to spare.

In one day the Magic Stat is programmed for the whole week and for the weeks to come. If you want a different schedule for weekends, you can individually program the thermostat for those days, too. "Big deal," you might think. "What's so great about that?" Read on.

You set most electronic setback thermostats to the time you want the furnace to go on in the morning, so when you wake up, your room is once again warm. But what if one morning it's bitter cold outside and the next morning it's much warmer? This means that setting your furnace to go on at the same time may, on one morning, leave you cold and on the next morning cause you to waste energy by warming up your house too soon.

By golly, the Magic Stat has everybody beat on this one too. Throughout the night it senses and computes the drop in temperature and the time it will take to get your room to your exact wake up temperature. So if you want to wake up at 7 AM to 70 degrees—that's the temperature you'll wake up to every time. Because it's a patented concept, no other thermostat has this feature. But wait. There is also a patent on the setting feature.

SIMPLE TO SET

To set the thermostat, you press just one button. A small LED light scans the temperature scale until you reach your desired temperature and then you release the button. You change the temperature naturally, throughout the day, up to six times. The unit responds and remembers that exact living pattern. The present temperature is displayed by a glowing red LED on the scale.

The system also computes the ideal length the furnace should stay on to keep the temperature within a range of plus or minus one and one-half degrees. A battery backup lets you keep your stored program in its memory so power outages as long as eight hours won't let your unit forget. And if something happens and your power is out for a few days, the unit will automatically maintain 68 degrees when the power is restored.

Quite frankly, we were so impressed with the unit, its ease of installation and setting plus its many energy-saving features, we seriously considered advertising it until we realized that our customers would probably not want to trust their future comfort to a product called Magic Stat. What if something went wrong with the unit? How substantial was this Magic Stat outfit? Remember, a thermostat is something you live with as long as you live in your home, and they're supposed to last ages. After all, your comfort depends on it.

Well, we did our homework, We found the company to be a sound, well-financed organization. They have been in business for several years, and they back their products with a three-year limited warranty. In addition, the company has a policy of buying back your unit in one year if you haven't saved its full cost in energy savings. We were satisfied with the company, the people, the product, its incredible features, the company's commitment to the product and above all, the energy savings.

We are so impressed now with the Magic Stat that we're going to make buying one irresistible. Buy one from us for only $79. Install it yourself in a few minutes or hire a handyman to install it.

Or order the new deluxe unit for $99 with the exact same features as the regular model, but with a beautiful new case.

Then enjoy the savings this next winter, Not only will you save up to 30% on your heating bills, but you're eligible for the 15% energy tax credit. Then if you're not absolutely in love with this product one year later, return it to JS&A. You'll get all your money back and you can reinstall your old thermostat.

REALIZE SAVINGS

But we're counting on a few things. First, you will realize an energy savings and a comfort that will far surpass what you are currently experiencing. Secondly, you probably will sleep better breathing cooler air yet wake up to just the right temperature.

Beauty is only skin deep and a name doesn't really mean that much. But we sure wish those guys at Magic Stat would have named their unit something more impressive. Maybe something like Twinkle Temp.

To order, credit card holders call toll free and ask for product by number below or send check plus $4 delivery for each unit ordered.

Magic Stat (0040F) **$79**
Deluxe Magic Stat (0041F) **99**

JS&A®
One JS&A Plaza
Northbrook, Illinois 60062
CALL TOLL FREE 800 323-6400
IL residents add 6% sales tax. ©JS&A Group, Inc.,1983









Success
is often measured
by how deeply
you're in the Black.

Johnnie Walker Black

12 YEAR OLD BLENDED SCOTCH WHISKY,
86 PROOF. BOTTLED IN SCOTLAND.
IMPORTED BY SOMERSET IMPORTERS, LTD., N.Y. © 1983

10157 ($32.50). The engineering handbook has come to have a generic look: a squat, heavy volume on thin paper tight with long tables of well-tried numerical data. The book bears the name but is no standard member of the genus. All the same it is a timely and useful book; routine specification and design of robots have simply not yet arrived.

This rounded compilation of matter of diverse source and nature fills a volume of unexceptional size and weight, neatly produced. It is an overview for readers with a broad understanding of engineering and economic topics; only here and there, particularly around its account of computer programs and robot control, does the level taken enter the domain of the specialist, say a would-be designer. The text begins with what amounts to a small book on robots in general, their configuration, sensors, tools, programs and control, safety and socioeconomic impact. Then we come to a look at the state of the art: a field guide to "current robotic systems" and a few pages of specifications and figures from each of the 28 U.S. manufacturers of robots, more than 80 species overall. Not at all provincial, the compiler then goes overseas; there is a review of industrial-robot manufacture in 10 countries, including a major review of the industry in Japan, although not in the U.S.S.R. The last 40 pages take a look at the robot and the future, by way of invited or cited statements from a wide range of expert and interested sources. The book is completed by a glossary, a bibliography and a brief summary of robot R&D organizations in the U.S.

Scale determines. There are industrial workers by the 100 million in the world. The robot population, by any plausible definition, is a thousandfold less. Growth is spectacular, but even by extrapolation from ebullient Japan it does not seem that robots can staff as much as a few percent of the work over the next few decades. The entry points for working robots seem understandable. Hard automation, say the familiar wonders of the light-bulb industry, is not robotic; its repetitive operations are already highly optimized for a well-defined product, and they are slow to change. Most small batch production, with hand assembly and inspection, changes so often and requires so little training of modestly paid human workers that it does not appeal to robots.

The robot's role lies in between, moving parts and assembling them up to a point, where the electrical-appliance and machine-tool industries offer plenty of niches. Other familiar tasks, notably welding and spray painting, and eventually work with hot metal, demand too much from human workers, who have the habit of breathing and require strict temperature control. A robot that works no faster than a human welder can triple output; the masked and overheated welder can work with that glaring arc only 30 percent of the work time.

Experience bears out this view: Nissan and Toyota and their counterparts everywhere are the biggest employers of robots. These automotive robots are playback devices; they learn on the job as a skilled worker leads the apprentice step by step through the hand task, typically making his points through the use of a hand pendant control. A Cincinnati-Milacron T-cubed robot, able to work a couple of shifts every day, sells for $63K, and it will store in a microprocessor the locations and sequence of more than 400 points taught it, moving between them by means of continuous-path servo, able to define positions absolutely and to reproduce them within about 1.3 millimeters. Its hydraulic drive will swing a couple of hundred pounds around six axes, and it has been certified as a spot welder, a parts-transfer-machine tender, a drill operator and more. The big one-armed deaf-blind fellow stands on a heavy three-foot base from which it can reach much of an eight-foot hemisphere. This is an experienced arm; it is state of the market, in stock today.

Control and program for robots transcend the ingenuities of the more familiar symbol-handling computers. The robot dwells in an untidy real universe, and the problems of mechanical oscillation, hydraulic drifts and urgent exceptions to the program commands (something fell!) are surely moving the designs toward smarter, more attentive devices. Breakaway wrists for tool arms sound crude; the T-cubed arm can drill far more precisely once it is given a smart drill, one with its own template, contact sensors and dedicated microcomputer. Robot locomotion is just ahead: more need for good robot sense and sensors. Ship welders, coal-face miners, underwater workers and window cleaners are ahead, and medium assembly seems wide open for the near term.

The robot future is unclear in detail, except for growth in number and sophistication. Two analysts, one for the United Automobile Workers, Daniel D. Luria, and one, James S. Albus of the National Bureau of Standards, offer clear-eyed views near and far. U.S. metalworking employment is bound to shrink. American white collars also will get bluer as offices increase in capital cost and require multiple shifts. Unless output grows with unexpected speed, serious dislocation is sure. The "bad" jobs should go to the robots first; the unions need a voice in technology choice; the market cannot determine the new structure for the U.S. economy. The workweek should dwindle; robot factories making robots, with an exponential hiding in there somewhere, loom on the far horizon.

SCIENCE/SCOPE

The space shuttle's new "eyes, ears, and voice" will revolutionize future missions, tests aboard the shuttle Challenger last June indicate. The integrated radar and communications system, also called the Ku band radar because of its operating frequency, uses an antenna dish at the front of the cargo bay. The system will let shuttle crews talk to Earth or transmit TV, high-speed data, and payload telemetry through NASA's tracking and data relay satellites. Crews now communicate with the ground less than 20% of the mission because the spaceship passes beyond the range of ground stations. With the Hughes Aircraft Company system, communications time will increase to over 90% of the mission.

Pioneer 10 is streaking into interstellar space with navigational help from its electronic imager. The spacecraft, which made history last June upon leaving the solar system, is using its imaging infrared photopolarimeter (IPP) to fix on the star Sirius. Pioneer 10 previously oriented itself with a sun sensor, but the sensor, now well beyond its design range, is reaching its limits of sensitivity 2.9 billion miles away. Pioneer 10 needs a reference point for spacecraft attitude control and interpretation of scientific data on solar wind. The IPP had been repeating various cruise-mode experiments since giving scientists their first close-up pictures of Jupiter and its four largest moons in late 1973. The IPP was built by the Santa Barbara Research Center, a Hughes subsidiary.

An advanced antenna farm designed with the aid of a computer will be carried into space by Intelsat VI communications satellites. The system will provide many different kinds of coverage -- beams transmitting to entire hemispheres, "global" beams, focused regional beams, and very narrow spot beams for broadcasting high-speed data. Hundreds of computer patterns were created to predict antenna performance. These studies led to the choice of transmit reflectors 3.2 meters in diameter instead of 4 meters. The larger size was rejected because it offered only slight improvement at the cost of being much heavier, larger, and more complex. Hughes heads an international team building Intelsat VI.

The new Landsat 4 spacecraft measures surface temperatures with an emitted thermal energy band not available on previous Earth resources satellites. The band, part of an instrument called a thematic mapper, adds a new dimension of data to that of bands that measure reflected sunlight. The new band can identify and map surface composition for geological studies and for mineral and petroleum exploration. It also can identify vegetation types and monitor their health. Hughes and its Santa Barbara Research Center subsidiary built the mapper for NASA.

"Diversification" characterizes Hughes Industrial Electronics Group, located in the Southern California communities of Carlsbad, Irvine, Newport Beach, Sylmar, and Torrance. Our facilities are like small companies with 500 to 2,500 employees, but each offers the resources and benefits of a multibillion-dollar company. Our diverse technologies include silicon and GaAs semiconductors, fiber optics, microwave and millimeter-wave communications, microprocessors, lasers, and solar cells. Send your resume to B.E. Price, Hughes Industrial Electronics Group, Dept. SE, P.O. Box 2999, Torrance, CA 90509. Equal opportunity employer.

Creating a new world with electronics

HUGHES

HUGHES AIRCRAFT COMPANY

For additional information please write to:
PO Box 11803, Marina del Rey, CA 90295

Established 1845 SCIENTIFIC AMERICAN October 1983 Volume 249 Number 4

Labor-intensive Agriculture

The $18-billion U.S. fruit and vegetable industry is increasingly reliant on illegal-immigrant labor. By postponing mechanization it is becoming vulnerable to cheaper produce from other countries

by Philip L. Martin

Agriculture is the largest single industry in the U.S., employing 4 percent of the labor force and generating 4 percent of the gross national product. The enterprise consists of 2.5 million farms that combine family labor and hired labor, land and water, seed and fertilizer, and energy and machinery to produce livestock and commodities worth nearly $150 billion per year. The U.S. food system is the envy of the world because it provides a wide variety of foods to consumers at low cost. The average family in the U.S. spends about 17 percent of its disposable income on food, whereas in Europe 25 percent is common and in developing countries the fraction can go as high as two-thirds.

Most farms in the U.S. are diversified operations where the farmer and his family plant and harvest several crops and tend livestock as well, all without any help from hired farm workers. Seventy percent of the nation's farm work is done by farmers; only the remainder is done by hired workers. Another part of the picture, however, is that 5 percent of the farms produce half of the nation's food and fiber, and it is these commercial farms that employ most of the hired workers. Many of the commercial farms that grow fruit, vegetables and nursery products—all labor-intensive crops—resemble nonfarm businesses, with hired managers and up to several thousand seasonal workers to do the farm work. These seasonal farm employees are increasingly illegal aliens or undocumented workers (people working in the U.S. without proper authorization). Agriculture's growing dependence on alien workers runs counter to U.S. law and policy, retards the pace of the labor-saving technological changes that have made the rest of U.S. agriculture a paradigm of efficiency and sends the wrong signal to farmers who face increased international competition. The continued availability of alien workers discourages farmers from making necessary labor and production changes, thereby increasing their vulnerability to foreign competition.

The annual farm-wage bill of $12 billion is divided into thirds: one-third is paid to workers on farms that produce fruit and vegetables, one-third to workers in field crops such as wheat, corn and cotton, and one-third to livestock workers. Livestock farms have few problems getting year-round salaried workers, and most field-crop farms can find as many seasonal equipment operators as they need. The major problems of farm labor arise in the $18-billion fruit, vegetable and nursery industry, which hires more than one million seasonal farm workers and usually pays them piecework harvest wages, for example 40 cents per bushel of apples.

Fruit and vegetable farms need workers for several weeks or months. Across the U.S. three times more farm workers are required for the harvests of September than are required for the winter tasks of January. Most fruit and vegetable farms grow only one or two commodities, thus increasing the ratio between peak and trough seasonal employment to as much as 20 : 1. The farm-labor market must match more than a million seasonal workers annually with farm jobs of short duration and must do so under conditions that include many workers who do not speak English, few incentives such as seniority and fringe benefits to hold workers to a particular farm, and perishable crops that may be lost if the harvest is delayed.

Most fruit and vegetable farms recruit hundreds and sometimes thousands of workers who are willing to be employed on one farm for a short time and then to look for another seasonal job. Traditionally the task of stringing together a series of farm-labor jobs has fallen on the worker. Since many seasonal farm workers speak little English, bilingual farm-labor contractors and other intermediaries such as employer associations and hiring halls have emerged to recruit, supervise and pay farm-worker crews that are moved from farm to farm. Seasonality, the contractor system and piecework wages have fostered a condition in which workers and jobs are interchangeable: farmers know the piecework cost of harvesting, and so they are indifferent about who picks their crops; since piecework wages are quite uniform, workers do not care who owns the field they are working in.

Debates over farm labor would leave less room for contradictory assertions if farm-labor statistics could provide an accurate profile of the workers. Unfortunately the statistics are incomplete and unreliable, with the result that each data series provides a glimpse of a scene that remains obscure in its details. This obscurity allows contradictory assertions to coexist: most seasonal farm workers are American students; most seasonal farm workers are illegal Mexicans, proving that Americans will not do farm work.

The profile that emerges from the statistics of the Department of Agriculture is that the typical hired farm worker is a 22-year-old white male college student who does one month of farm work during the summer. Since even department spokesmen testify in Congress that an immigration reform reducing the availability of alien farm workers would disrupt the fruit and vegetable industry, one can infer that the department's sta-

tistics systematically undercount seasonal fruit and vegetable workers.

There are two major sources of farm-labor data. Each July the Department of Agriculture asks farmers how many workers are employed, what they are paid and what fringe benefits they receive. Every other December the department pays the Bureau of the Census to attach farm-labor questions to the monthly Current Population Survey, which is the source of the national figures on employment and unemployment. About 1,500 of the households in the survey include at least one farm worker, and the responses from this group serve to generate a profile of farm workers.

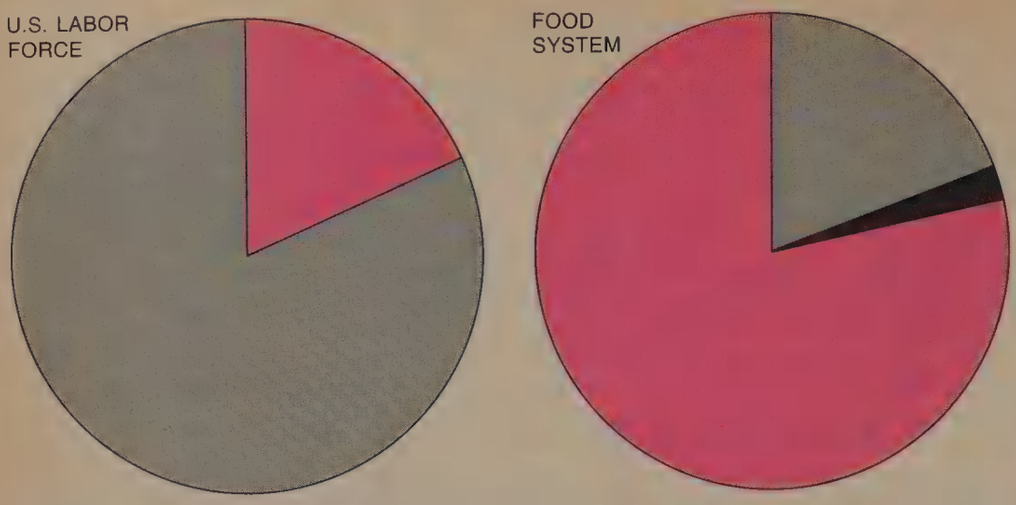
The July survey estimates the employment of farm workers, that is, the number employed during the week of the survey. The December survey estimates the hired farm work force, that is, the number of people who did farm work for wages sometime during the year. An industry's total work force is always larger than employment because of turnover among workers.

The Department of Agriculture reports that the average employment of farm workers is about 1.3 million throughout the year and that the total farm work force is about 2.5 million. The implication is that a farmer must normally hire two workers during the year to keep one job slot filled. Since the workers on many livestock and field-crop farms are employed year-round, a reasonable assumption is that the turnover is highest on the fruit and vegetable farms hiring seasonal workers. Indeed, some farmers have reported hiring 200 workers in one month to maintain a 20-person harvest crew; at an annual rate this would be 120 workers to keep one job slot filled for 12 months.

Such an extreme turnover precludes a clear picture of who does farm work. The employment survey of July, 1982, reported that 1.8 million workers were employed on 475,000 farms, earning an average of $3.96 per hour. Farmers reported that they provided fringe



MANUAL LABORERS pick green peppers on a farm in Gilroy, Calif. Like most fruits and vegetables intended to be sold fresh, the green peppers must be picked by hand; mechanization extends mostly to fruits and vegetables that are processed before they are sold.



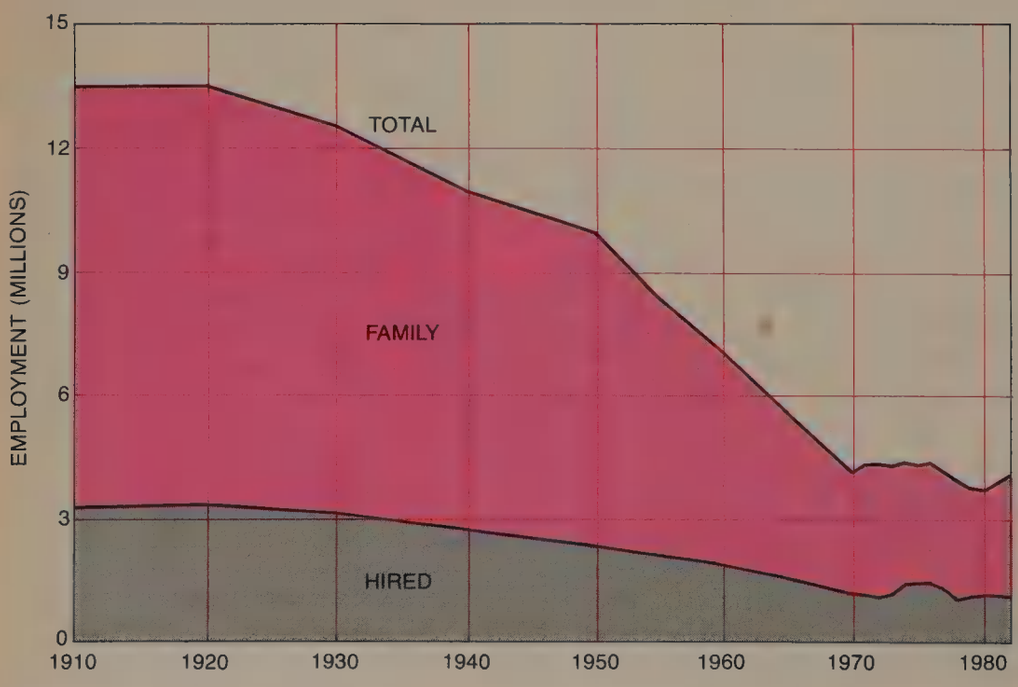
COMPONENTS OF EMPLOYMENT are analyzed. Of 95.9 million people employed in the U.S. in 1980 some 17.6 million, or 18.3 percent, were in the food system (*color*). That group is further analyzed in the chart at the right. Food processing and distribution accounted for 13.8 million workers, or 78.5 percent of the food-system work force, farming (*gray*) for 3.3 million (about 19 percent) and farm inputs such as seed and fertilizer (*black*) for 411,000 (2.3 percent).

benefits such as housing, meals and transportation to 45 percent of their employees, although the actual value of these benefits ranged from a carton of the crop being harvested to free room and board. An "average" farm worker employed during the week of July 11–17, 1982, worked 35.3 hours at $3.96 per hour, a weekly wage of $139.79. (In 1982 the average wage of production workers in industry was $7.89 per hour.)

The work-force survey in December, 1981, estimated that 2.5 million people did farm work for wages sometime during the year. They were (according to the survey) mostly white (73 percent), male (77 percent), young (55 percent under age 25) and poorly educated (median 11 years). Most respondents worked for only a short time in agriculture, 73 percent of them for fewer than 150 days. The people who worked more than that constituted only 26 percent of the farm work force but did three-fourths of all the farm work done by hired hands. Migrants (people who cross county or state lines and stay away from home at least overnight to do farm work) made up only 5 percent of the total work force. According to this survey, the average farm worker in 1981 earned $27 per day for 98 days of farm work, or $2,659 from agriculture.

The picture derived from these statistics on the work force is distorted because the "average" characteristics are determined by the large number of students who do a few days or weeks of farm work in the summer and are easy to find in December. Indeed, the data indicate that 34 percent of the "farm workers" were primarily students, considerably more than the 28 percent whose primary activity was hired farm work.

Farm-labor data are the source of much frustration to farmers, farmworker representatives and policymakers. Notwithstanding the evidence that the seasonal farm work force in the Southwest is largely Mexican, the data from the Department of Agriculture would have it that young, white American men predominate. The people who are skeptical of these statistics assert that a survey conducted every other December will find students and will miss the many Mexicans who leave the U.S. during the winter months. Indeed, a survey in 1981 of 472 seasonal-farm-worker families in the Central Valley of California found that more than 90 percent of the workers were Mexicans and Mexican-Americans. Most representatives of farmers, farm workers and government agencies say the work force of from 700,000 to one million in the fruit and vegetable industry consists largely of Mexican nationals, many of whom are working in the U.S. illegally.



TREND IN FARM WORK between workers who are members of the family and workers who are hired shows that although employment has declined in both categories, the fraction of hired workers was significantly higher in 1980 (35 percent) than it was in 1910 (25 percent).

Four kinds of aliens do seasonal farm work: immigrants, border commuters, H-2 workers (named for the section of the Immigration and Naturalization Act of 1952 that authorizes their entry) and illegal residents. Immigrants are aliens admitted to live permanently in the U.S. Although most of them settle in urban areas, from 10 to 15 percent find jobs in rural areas. Border commuters are immigrants who maintain a home in Mexico or Canada and commute daily or seasonally to jobs in the U.S. There are about 60,000 of them, many working on farms.

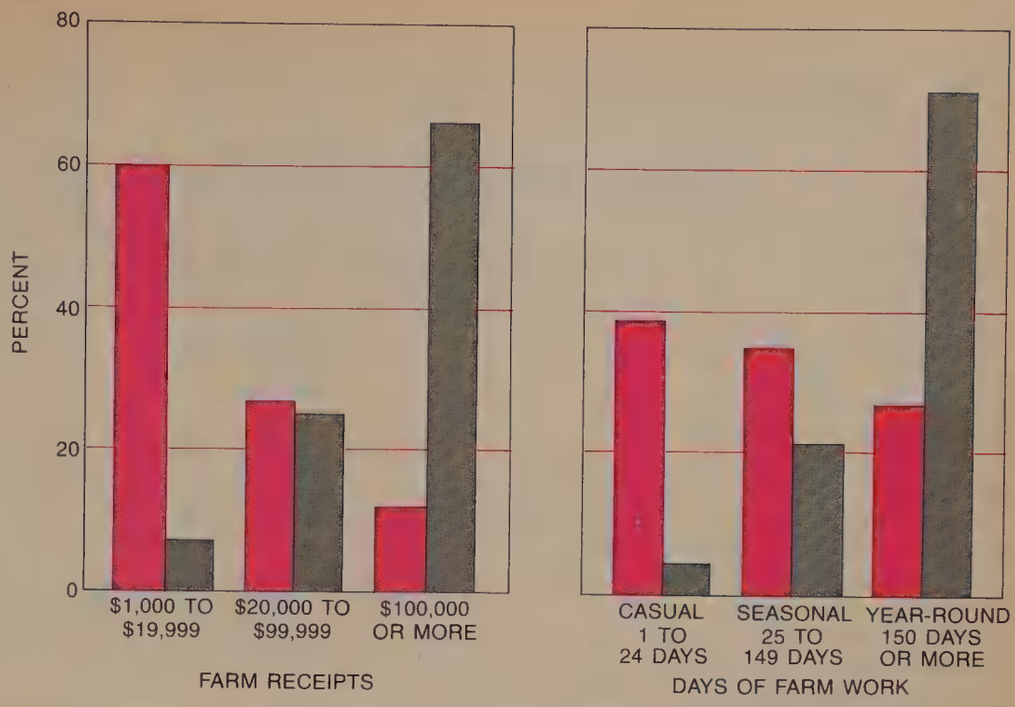
The H-2 program was enacted in 1952, when the separate bracero (manual laborer) program was almost at its peak. Under the H-2 section aliens are admitted to the U.S. for temporary jobs if unemployed U.S. citizens are not available and if the employment of alien workers will not adversely affect similarly employed Americans. The agricultural part of the program admits about 15,000 aliens annually for farm jobs that last from two to six months. H-2 workers cut sugarcane in Florida, pick apples in the mid-Atlantic states, harvest citrus fruit in Arizona and herd sheep in the West. Although they are only a small fraction of the total seasonal farm work force, they are predominant in the areas where they work and in the tasks they do.

The largest alien work force in agriculture consists of illegal aliens or undocumented workers, who come primarily from Mexico and other Latin-American countries. The bracero program, which lasted for 22 years, admitted almost five million Mexican farm workers to the U.S.; after it was terminated in 1964 farmers obtained immigrant

status for many of their best workers. The former braceros who settled in the U.S. and those who continued to commute across the border developed efficient networks that informed relatives and friends about jobs in the U.S. and provided advice on how to cross the border illegally. The networks are still in place and can deliver additional workers to farmers in the U.S. on short notice. A new network is created whenever a venturesome or desperate worker leaves another one of Mexico's 75,000 villages and finds work in the U.S.

The Simpson-Mazzoli bill on immigration reform and control that is now before Congress would (1) make it a Federal crime to knowingly hire or employ an illegal alien, (2) grant amnesty to certain aliens living illegally in the U.S. and (3) modify the H-2 regulations to make it easier for employers to get alien farm workers. The assumptions behind these three related changes are that sanctions will reduce the influx of illegal alien workers, that amnesty will encourage some people to leave agriculture for nonfarm jobs and that a modified H-2 program will be necessary to prevent "disruptions" in agriculture.

The basic administrative question in the H-2 program is simple: Where does the duty of the farmer to hire an American work force stop and the obligation of the Federal Government to open the border start? Clearly if farmers have to scour the U.S. looking for workers and must offer high wages, transportation, adequate housing, inexpensive food and other amenities, they will find more American workers and will need fewer foreign ones. They will also have a greater incentive to mechanize. On the other hand, if farmers have few obligations in recruitment, wages and housing, they will find it easier and cheaper to simply hire aliens. The H-2 program is controversial because it is trying to strike a balance between conflicting goals: protecting American farm workers and ensuring a plentiful supply of low-cost farm labor.



ECONOMICS OF FARMING are indicated in these related charts of cash receipts (*left*) and hired workers (*right*). In the chart of farm receipts the colored bars show the fraction of all farms in each category and the gray bars show how the farms in each category shared in the cash receipts. The biggest share of the income went to the largest farms. Similarly, in the chart of hired workers the colored bars represent the percentage of all such workers in each category, the gray bars the percentage of work they did. Year-round employees did most of the work.

The debate over the need for alien workers in fruit and vegetable farming diverts attention from an unsettling shadow on the horizon: competition from other nations. The automobile industry faced a similar juncture in the late 1960's. In an affluent and health-conscious society the demand for fruits and vegetables is expanding, and farmers envision continued expansion and profits if the Federal Government does not enact costly new labor regulations. Just as the automobile industry waged largely successful battles against air bags and environmental controls that diverted attention from increasing foreign competition, so fruit and vegetable farmers are fighting to continue using seasonal alien farm workers in order to compete with the fruits and vegetables other nations are beginning to export in quantity. Cheap labor benefits agriculture in the short run, but it also helps to blind farmers to the technological changes they will have to make in order to compete with foreign producers who have access to even cheaper labor.

Products such as citrus fruits, strawberries and tomatoes require relatively little land but a great deal of labor. The cost of hand harvesting is 20 percent of the price the farmer gets for oranges and lemons and up to 40 percent of the price for lettuce, strawberries and tomatoes. The wages of farm workers in the U.S., however, are five times higher than they are in Greece and 10 times higher than they are in Mexico. As other nations expand their labor-intensive agriculture, the already loud complaints by American farmers about Brazilian oranges, Greek raisins, Mexican tomatoes and Italian wine will intensify.

The increasing dependence of American fruit and vegetable farmers on alien workers confronts policymakers with two options: (1) preserve the status quo by approving an open-ended temporary-worker program or (2) encourage the fruit and vegetable industry to mechanize in order to limit its need for alien farm workers. If the U.S. gradually reduced the supply of alien farm workers, the production of labor-intensive commodities that cannot easily be mechanized could be shifted abroad by removing import barriers, thereby increasing employment in Mexico and the Caribbean basin and decreasing incentives to migrate illegally to the U.S.

Debates on farm labor often fail to recognize these linkages of immigration, trade and technology. Instead many fruit and vegetable farmers believe the optimal strategy is to preserve the status quo with an open-ended temporary-worker program. They recognize that mechanization is inevitable, but they want to decide when and how machinery will replace workers instead of being confronted with the sudden elimination of traditional labor supplies. Organizations of farm workers oppose a temporary-worker program because they believe aliens selected from a huge pool of foreign labor will always be preferred to Americans, who are often the castoffs of other labor markets. These organizations also oppose mechanization that displaces farm workers.

Machines have taken the place of seasonal workers in the fruit and vegetable industry in an uneven pattern since research on this kind of mechanization was begun in the 1940's. Most of the root vegetables such as potatoes are dug mechanically, some of the leafy vegetables such as spinach and processing cabbage are lifted and cut by machine, and a few of the field crops such as pickling cucumbers and processing squash can be machine-harvested. Crops that are intended to be sold fresh, however, still call for hand labor; for example, most carrots and green onions are still pulled and bunched by hand, and iceberg lettuce, broccoli and celery are hand-cut and hand-packed. Strawberries, melons and most fresh-market tomatoes are hand-picked.

Mechanization slowed in the 1970's

because workers were readily available. Nevertheless, harvesting machines are still being adopted by some farmers. Grapes that are processed into wine can be mechanically harvested by a machine operating over the row to shake the grapes loose and catch them on conveyor belts. This machine is being modified to harvest the raisin grapes that are now picked by hand. Most nut crops such as almonds, pecans and walnuts are harvested by a hydraulic arm that grasps the tree trunk and shakes the nuts to the ground, where they are swept up by another machine. A tree shaker with a padded catcher can harvest prune plums, cling peaches, apples and tart cherries, but the machine damages the fruit so much that the crop must be processed soon after harvesting.

Several of the crops that are now picked by hand could be harvested mechanically if workers were not available; among them are peaches, lettuce and tomatoes. The mechanization of many others will require the coordinated efforts of plant scientists, engineers, growers and processors. Trees and plants that yield uniformly ripening crops able to withstand machine harvesting could be developed. Growers would have to prune or plant them in such a way as to facilitate mechanical harvesting. Engineers could improve the optical and laser sorting devices that can sort the crop in the field, thereby cutting costs, but processors would have to schedule their work precisely so that the harvested and sorted produce would not deteriorate. Equipment manufacturers and seed companies could work with growers to diffuse new technology on the one hand and new plant varieties on the other.

The most plausible coordinator of this systems approach to mechanization is government, primarily through the established land-grant universities. These nonprofit institutions do the basic plant and engineering research that growers, processors and equipment manufacturers do not undertake. Moreover, the academic people who are engaged in this work have enough influence on both growers and processors to encourage changes in the practices that impede mechanization.

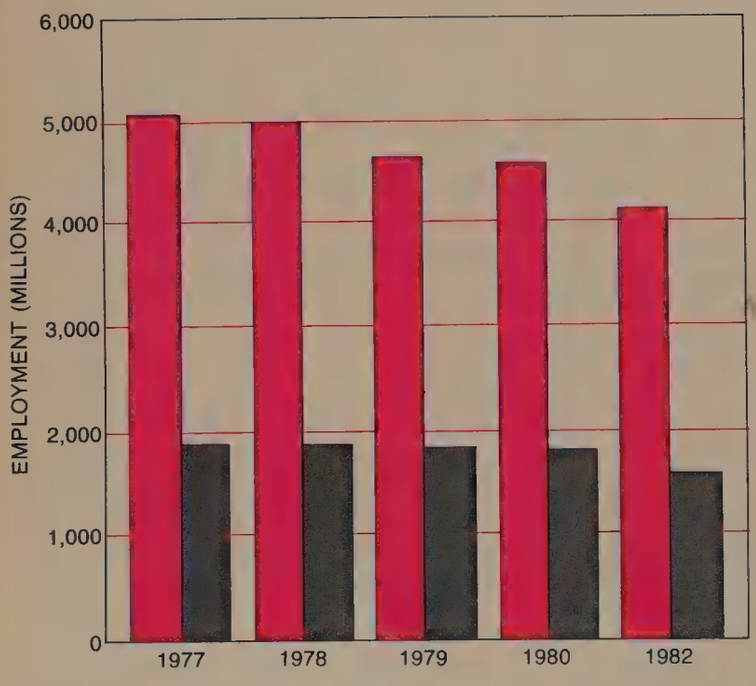
An increased role by the universities will be controversial, particularly in California, where the University of California is defending itself in a lawsuit brought by representatives of farm workers. The suit charges that the university's research on mechanization is biased in the direction of helping the operators of large farms and undermining farm-worker unions. Universities can head off some of the controversy by making clear the importance of mechanization in preserving a profitable fruit and vegetable industry and by taking pains to avoid conflicts of interest in research programs.

Public policy is trying to discourage agriculture from using aliens by enacting laws and making regulations forcing farmers to meet certain standards of recruitment, wages and working conditions. A better way to reduce dependence on alien workers is to link immigration reform to a plan for restructuring the fruit and vegetable industry. One way to forge such a link is with an H-2 trust, which would levy a tax on the wages earned by alien farm workers and would apply the funds to restructure the production of commodities that depend on alien workers.

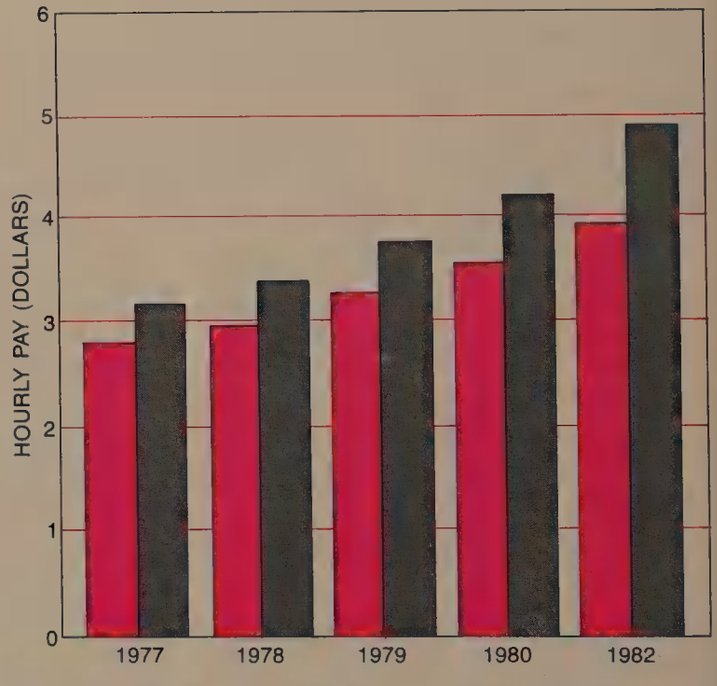
Such a trust could be established for each commodity whose growers asked for alien workers. The amount of money available to each trust would be determined by the level of the payroll tax and the extent of the commodity's dependence on alien labor. One can arrive at an estimate of what the level of taxation might be by taking into account the fact that farmers do not now have to pay the Social Security tax of 6.7 percent or the Federal unemployment-insurance tax of .8 percent on the wages of H-2 workers. In addition most states exempt those wages from the state unemployment-insurance tax (ranging from 3 to 6 percent) that farmers must pay on the wages earned by most of the farm workers who are U.S. citizens. These various exemptions suggest an H-2 wage tax of at least 10 percent. If a modified H-2 program admitted 300,000 alien farm workers annually and the average worker earned $5,000 during the year, the H-2 trusts would collect 10 percent of $1.5 billion, or $150 million per year.

Farmers are familiar with the concept of assessments to support commodity-related research and promotion. Since H-2 trusts could be established on a commodity-by-commodity basis, administrators familiar with the problems associated with the particular commodity could fine-tune the programs for reducing dependence on alien labor, choosing from a variety of options including research on mechanization, the training of workers and changes in processing. A commodity-specific H-2 trust would create both a direct link between a commodity's dependence on alien workers and the means to end that dependence.

The U.S. Select Commission on Im-



EMPLOYMENT AND WAGES in agriculture are characterized. In the chart at the left the colored bars show the number of farm workers in each year, the gray bars the number who were hired workers.



In the chart at the right the colored bars show wages paid at an hourly rate, the gray bars piecework wages. The data are based on a survey made each July by the United States Department of Agriculture.

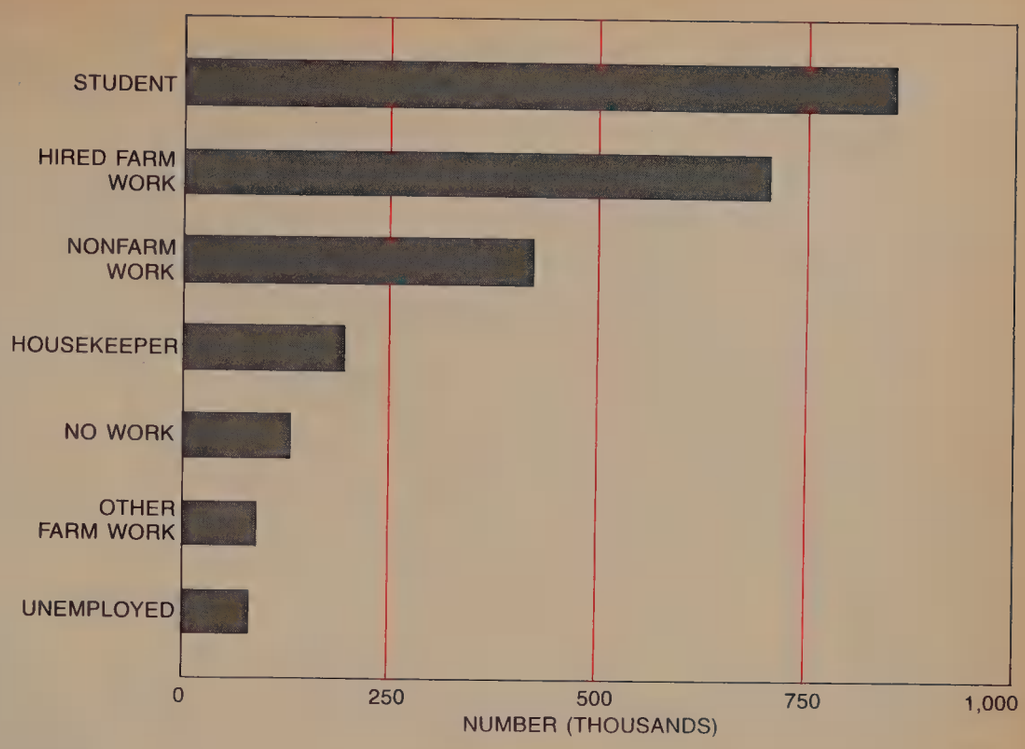
migration and Refugee Policy, which studied the subject from 1979 to 1981, recommended that no industry depend indefinitely on alien workers. The commission argued that such workers constitute a subsidy to employers unable or unwilling to attract American workers, to mechanize or to produce abroad, and it could find no justification for such a subsidy just to maintain a given industry in the U.S. If one accepts this reasoning, H-2 trusts would provide a way of emphasizing the transitional nature of alien farm workers in fruit and vegetable agriculture.

Mechanization is not the only way to reduce agriculture's dependence on alien workers. Intermediate back-saving technologies such as conveyor belts to eliminate heavy field bags and hydraulic lifts to replace ladders and fruit bags would help to encourage women and older workers to harvest crops. Experienced farm workers might prolong their average of from five to 10 years in harvest work if more farmers and employment associations offered a series of jobs that reduced unemployment.

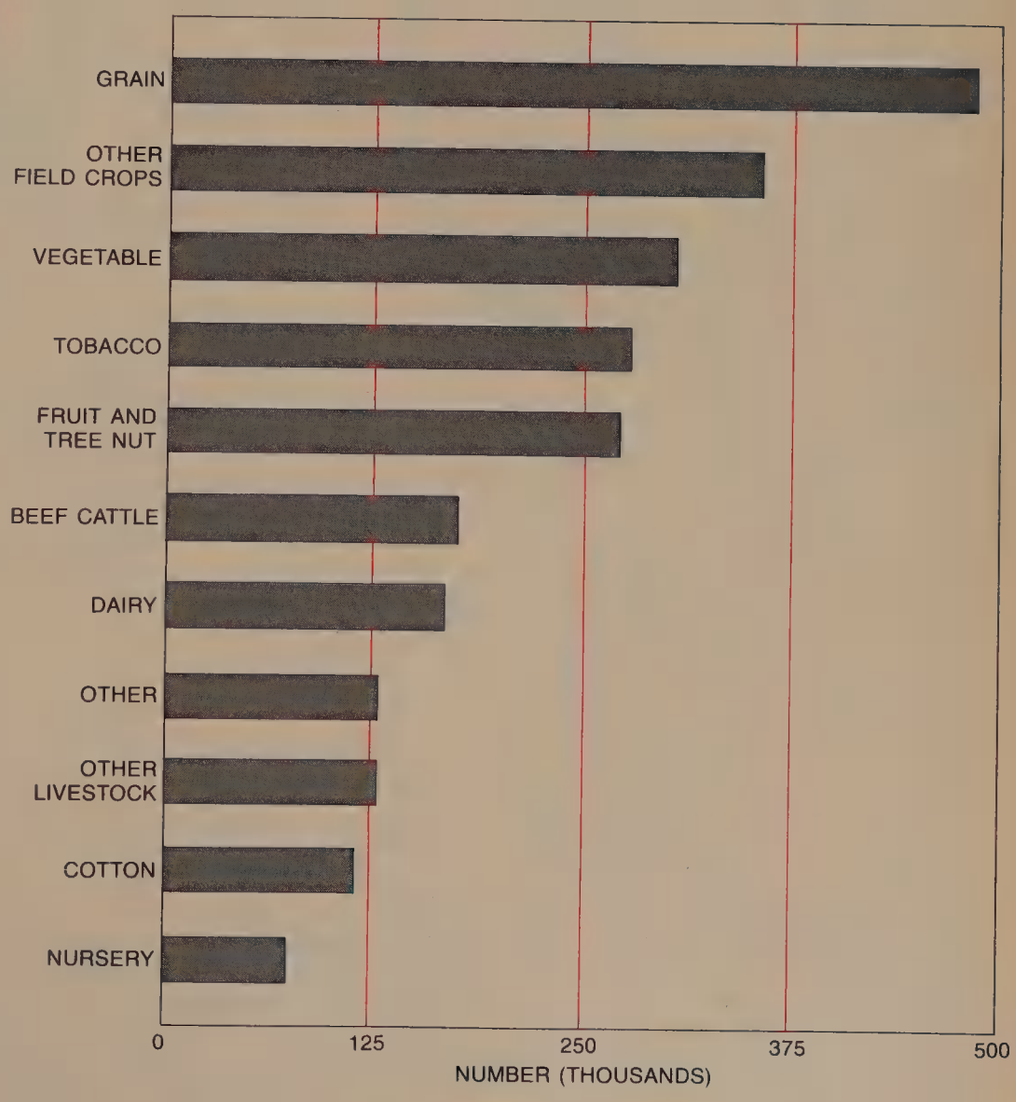
The continued dependence of the $18-billion fruit and vegetable industry in the U.S. on alien farm workers for handwork spells disaster in the long run. The industry has two choices: it can move toward mechanization, meanwhile trying to improve conditions for handworkers by adopting modern personnel policies and installing equipment that facilitates handwork, or it can obtain access to a rotating pool of alien farm workers with another program of the bracero type. The second option would imply an indefinite dependence on an alien work force.

Few U.S. citizens will become seasonal hand-harvest workers at the prevailing wages or even at higher wages because workers with options reject the uncertainties of the seasonal farm-labor market. The U.S. will have an American work force in fruit and vegetable agriculture when that industry is mechanized. The farmers who grow wheat, corn and soybeans can find Americans to operate their machinery, and the fruit and vegetable farmers who have mechanized are having the same experience.

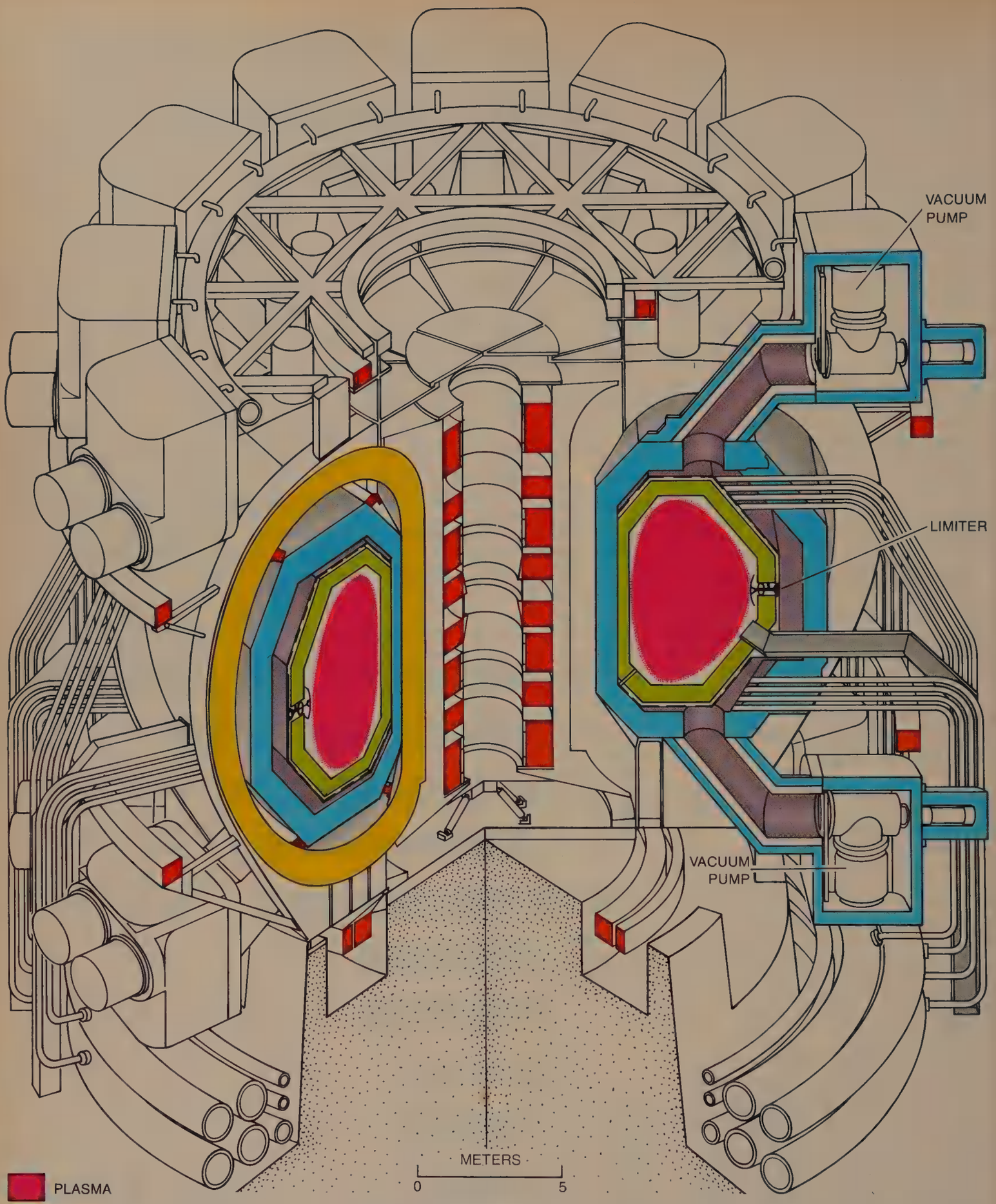
Mechanization is one answer to the problems of productivity, labor and immigration in fruit and vegetable agriculture. Without mechanization the U.S. must both accept an isolated, alien-dominated labor force for seasonal handwork and erect trade barriers to keep out produce grown abroad at even lower wages. If farmers successfully oppose the immigration reforms that could begin to alter this picture, they may win the short-run battle over labor but will lose the long-run war for survival in the increasingly competitive international fruit and vegetable economy.



PRIMARY ACTIVITY of hired farm workers in 1981 was in many cases something other than farm work. The data are based on a survey made every other December by the Bureau of the Census. They have been challenged by representatives of farm workers as overemphasizing the role of students, who can readily be found in December, and as underemphasizing the role of alien workers, mainly from Mexico, who do not all remain in the U.S. during the winter.



KIND OF FARM WORK done by hired workers in 1981 is charted according to the number of workers in each type of crop production. In the fruit and vegetable sectors many of them are aliens doing handwork; most of the hired workers for other types of farm are U.S. citizens.



STARFIRE FUSION REACTOR emerged on paper from a two-year design study led by the Argonne National Laboratory; it employs magnetic fields to confine a fusion plasma (a gas of ionized deuterium and tritium, the heavy isotopes of hydrogen) inside a toroidal chamber. Two cuts through the torus are shown. The left cut passes through one of the 12 superconducting magnets that form hoops around the torus. Other superconducting magnets are concentric with the torus. The right cut shows the equipment positioned between the hooping magnets. Vacuum pumps remove the plasma, which makes contact only with a "limiter" protruding through the "first wall" of the reactor. Pipes carry coolant to the first wall and a "blanket." Ducts guide radio-frequency waves through the blanket. The waves help to drive a current in the plasma and also help to heat the plasma to a temperature of tens of millions of degrees. In each fusion reaction deuterium and tritium combine to produce helium, a neutron and energy.

The Engineering of Magnetic Fusion Reactors

Design projects under way and experimental reactors now being built will test the practicality of schemes for generating power from the thermonuclear fusion of ions trapped by magnetic fields

by Robert W. Conn

The quest to tap the energy of nuclear fusion by employing magnetic fields to confine an ultrahot plasma (an electrically neutral gas consisting of charged particles) and generate electric power has been in progress for more than three decades. It is now approaching fruition. At a construction cost of $314 million the U.S. has built the torus-shaped Tokamak Fusion Test Reactor (TFTR), which began operating last December at the Plasma Physics Laboratory in Princeton, N.J. The European Economic Community is building the Joint European Torus, also called the JET tokamak, at the Culham Laboratory near Oxford. Initial operation has recently begun. A comparable device, the JT-60 tokamak, is being built in Japan; operation is expected in 1985. Together these three reactors represent a worldwide investment of more than $2 billion. The experimental campaigns planned at TFTR and JET will culminate in 1986 with the "burning" of a deuterium-tritium mixture. (Deuterium and tritium are the heavy isotopes of hydrogen: H-2 and H-3. They are also the fuel that a commercial power-generating reactor will consume.) The JT-60 will consume only deuterium and the abundant, light isotope of hydrogen: H-1.

An alternative to the toroidal confinement scheme of a tokamak is also being explored. In the U.S. the Mirror Fusion Test Facility B (MFTF-B) is under construction at the Lawrence Livermore National Laboratory in California. It is a tandem mirror, a device in which a plasma is confined by magnetic and electrostatic barriers at each end of a linear sequence of magnets. At a potential construction cost of $209 million, it is scheduled for operation in 1986. The aim of MFTF-B is to confine a deuterium plasma under conditions approximating those required for a commercial tandem-mirror reactor. In this sense the MFTF-B will test the scientific feasibility of the tandem-mirror approach. In order to minimize its consumption of electricity the MFTF-B is being constructed with superconducting magnets. An important set of coils, a pair of *C*-shaped magnets called a yin-yang configuration, has already been built and successfully tested. Two large tokamaks, the *T*-15 in the U.S.S.R., scheduled for completion by 1986, and the TORE-SUPRA in France, scheduled for completion in 1988, will also have superconducting magnets.

The decisions to construct these major experimental machines were taken in the 1970's in recognition of three key factors. First, there had been (and continues to be) a series of remarkable achievements in magnetic-confinement research. Second, the vulnerability of the industrialized world to interruptions in its energy supply had been amply demonstrated by events such as the Organization of Petroleum Exporting Countries (OPEC) embargo of 1973. Third, there arose an awareness of possible limits to both the amount and the use of fossil fuels. In the 1980's, then, the fusion experiments will get under way, and if they perform as expected, the scientific feasibility of a magnetic fusion reactor will be established. Still, a question will remain. Is magnetic fusion practical in terms of engineering, economics and safety?

The fusion reaction between deuterium and tritium requires that a stringent condition be met: the electrostatic repulsion between the deuterium and tritium nuclei (deuterons and tritons), which each have one unit of positive electric charge, must be overcome by making the relative speed of the two equivalent to an energy of about 100,000 electron volts (100 keV). A gas of deuterium and tritium has a broad distribution of energies; hence an average of 5 keV is enough for the gas to sustain its thermonuclear burning. That corresponds, however, to a temperature of 58 million degrees Celsius. At such a temperature electrons are not bound to nuclei, giving rise to the fusion plasma.

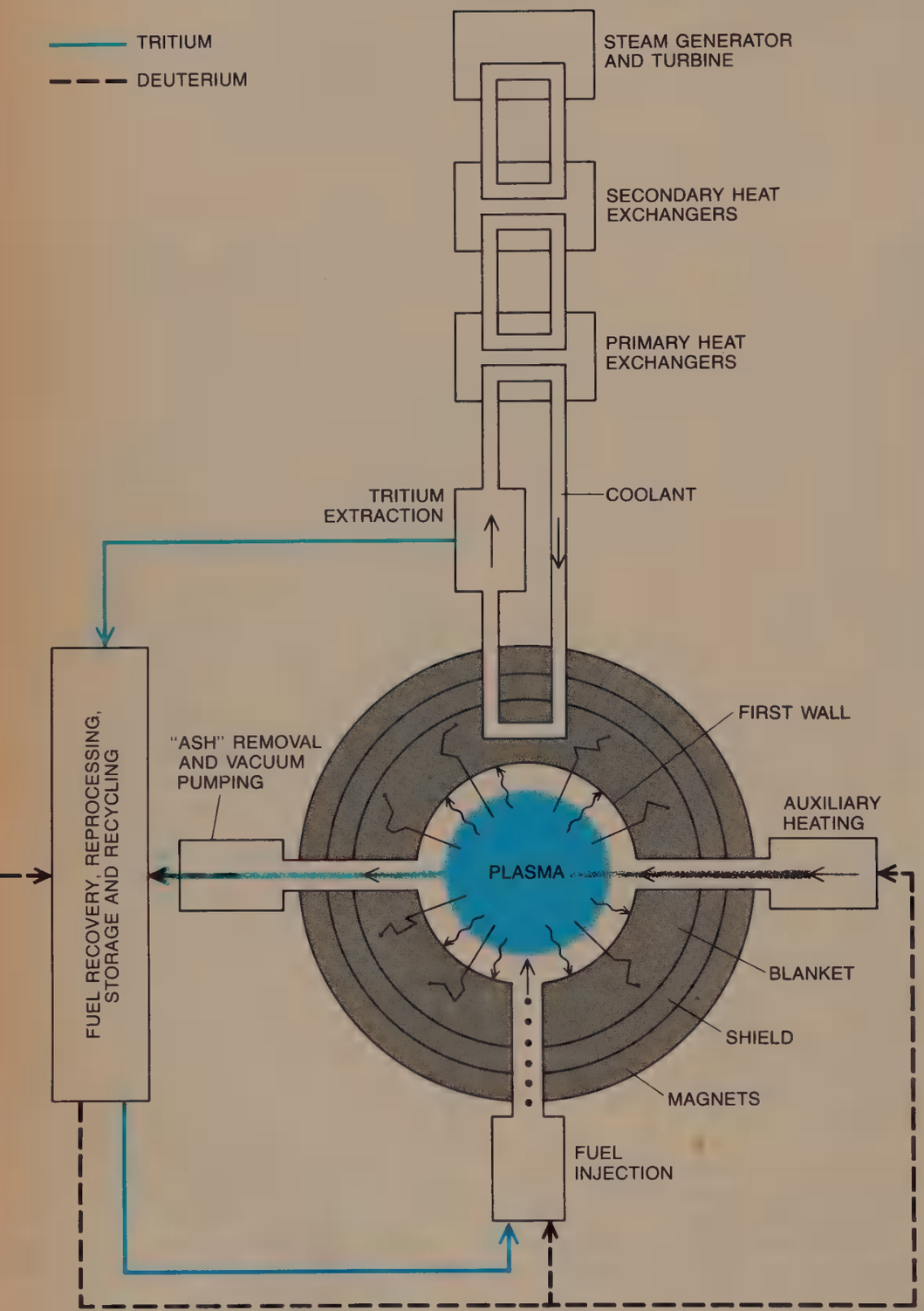
Each fusion reaction produces an energetic helium-4 nucleus (an alpha particle) and a high-velocity neutron. Because the neutron is electrically uncharged, it escapes from the plasma. The alpha particle does not escape; its positive charge of two units means it is trapped by the same magnetic field that serves to confine the plasma. With an energy of 3,520 keV it is much more energetic than the particles in the plasma, and so as it is slowed by collisions it gives its excess energy to the deuterons and tritons, thereby helping to maintain their temperature.

The necessary condition for the plasma to sustain its own thermonuclear burning—in a word, the requirement for ignition—is that the "confinement parameter," the product of plasma density (in particles per cubic centimeter) and average energy-leakage time (in seconds), exceed 3×10^{14}. If the confinement parameter is less than that value, the plasma's temperature cannot be maintained without auxiliary heating. The plasma can nonetheless yield energy. In effect it can act as a high-gain amplifier of the auxiliary power. The measure of the gain is the Q value, where Q is the ratio of the fusion power produced to the power injected to sustain the plasma temperature. At ignition Q becomes infinite. A reactor with finite Q is termed a driven machine because its fusion rate responds to changes in the auxiliary power input.

How should the plasma be confined? The broadest classification of magnetic confinement schemes depends on whether the magnetic lines of force are closed or open. Field lines that define a torus-shaped volume are the basic example of the former. Each charged particle in the plasma is locked to one of the lines. Specifically, it gyrates around the line but is free to move along it, thus

describing a corkscrew trajectory. Since the line meets itself, the particle is trapped. In the early approach called the stellarator the lines are produced entirely by coils outside the plasma. The plasma, once it ignites, can burn without external infusion of energy, and so the Q is infinite. In another approach, the tokamak, the lines are produced in part by external coils and in part by an electric current induced to flow in the plasma itself. In essence some of the external coils employed in a stellarator are replaced by a set of coils that act as the primary windings of a direct-current transformer. The plasma acts as a one-turn secondary winding.

Unfortunately a transformer cannot drive current in the secondary in one direction continuously because the current in the primary would then have to increase indefinitely. The transformer must thus be reset, and so the thermonuclear "burn" must be allowed to terminate periodically. Present estimates call for tokamak reactors to produce energy in pulses lasting for about 1,000 seconds with pauses of from 30 to 60 seconds. An energy-storage system would ensure a continuous supply of power to the turbines generating electricity in a commercial reactor installation, but the pauses would clearly impose design limitations on reactor components because of an increase in cyclic fatigue.



ENGINEERING PRINCIPLES common to all magnetic fusion reactors are diagrammed. A magnetic field must confine the fusion plasma; an auxiliary heating system must help to raise its temperature; a fuel-recycling system must keep it pure and well supplied with thermonuclear fuel. The heat the plasma releases must be withstood by the first wall of the reactor. The neutrons the plasma releases must penetrate into the blanket, where the energy the neutrons deposit must be transferred (in the form of heat) to a coolant. In turn the coolant can generate steam, the steam can drive turbines and the turbines can generate electricity. Nuclear reactions in the blanket must also "breed" tritium, which is radioactive and is extremely rare in nature.

Other ways to drive a tokamak's plasma current are therefore being investigated. If they are successful, they will enable tokamaks to operate steadily, or at least in very long pulses, on the order of one day. The most promising idea is to launch unidirectional radio-frequency (R.F.) waves into the plasma at a frequency of about three billion hertz. The waves' momentum can be transferred to the electrons, pushing them along the magnetic field and thereby driving electric current. Experiments at the Massachusetts Institute of Technology and at Princeton University have demonstrated the process. The key question is whether it works efficiently at the density of a fusion plasma. Theoretical estimates give values of from 10 to 20 for the Q of a reactor with R.F. current drive. A Q of 20 is about as low as one can accept for a commercial tokamak reactor.

In the second broad class of magnetic confinement schemes the field lines are open. The most basic configuration is a single-cell magnetic mirror. Here a set of magnets produces a field that has two peaks in its strength. Each peak is a "bottleneck" region, where the lines of force are close together. Consider a charged particle locked to one of the lines somewhere between the two peaks. As the particle approaches either peak it must continue to conserve energy and angular momentum; thus its energy (and motion) parallel to the field is converted into energy (and motion) perpendicular to the field. The motion parallel to the field can actually be reversed, so that the particle is trapped between the peaks.

A single-cell magnetic mirror is not a perfect container: particles colliding and scattering off one another can leak through the throat of a bottleneck. Electrons scatter much more rapidly than ions and therefore leak out faster. As a result the remaining plasma develops a positive electrostatic potential, which reduces the further flow of electrons just enough so that the currents of electrons and ions match. To sustain the density of ions, beams of neutral atoms at an energy of some 200 keV could be injected into the plasma. The neutral atoms would pass through the magnetic field unaffected and be ionized and trapped

when they collided with the plasma's electrons. The Q of a single-cell magnetic mirror sustained in this way is calculated, however, to be only about 1.

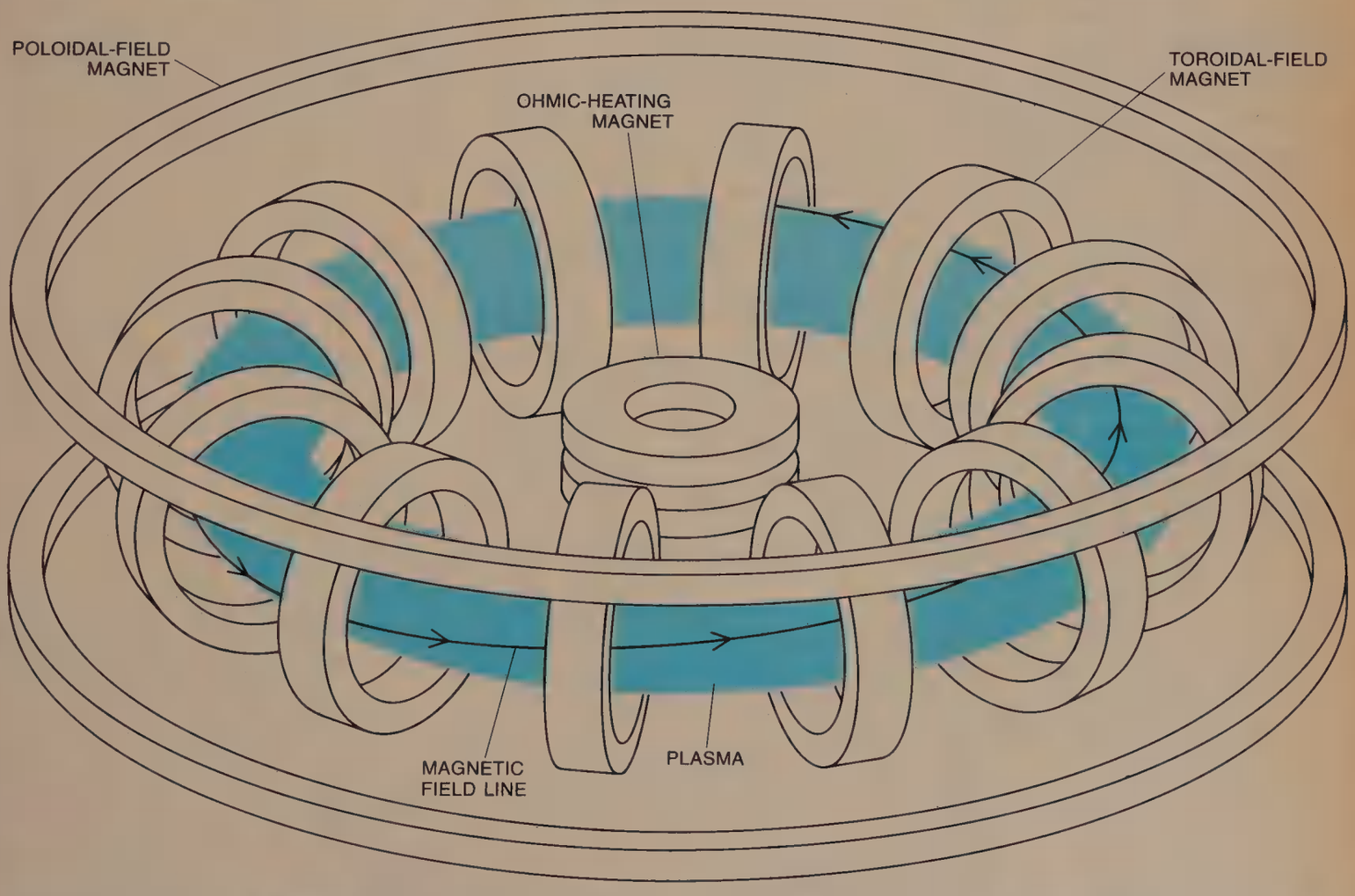
Alternatively two mirror cells can be connected magnetically to form a tandem mirror. The peak of electrostatic potential in each cell now acts as a "plug" to trap ions electrostatically in a central valley of potential between the mirror cells. Meanwhile a smaller but deeper valley inside each mirror cell serves to isolate the electrons in the cells from those in the central valley. This barrier allows the mirror-cell electrons to be heated to a high energy, thus creating the mirror-cell potentials quite efficiently. The electrons are heated by R.F. waves whose frequency, typically from 30 to 100 billion hertz, matches the electron cyclotron frequency (the frequency at which electrons gyrate around magnetic field lines in the plasma). The heating of end-cell electrons and the plugging of central-cell ions have both been demonstrated recently at Livermore.

Energy is required continuously to sustain the mirror-cell plasmas in a tandem mirror. Estimates are that its Q is between 10 and 30. It is important to note, however, that the energy of charged particles that do leak from the ends of the reactor can be recovered by directing the particles into devices that are in essence the reverse of particle accelerators. The production of electricity by such "direct convertors" (to supplement the power produced by the fusion reactions) can be 50 percent more efficient than the production of electricity by heating water and directing steam into a turbine. The estimates suggest that the minimum Q required of a practical tandem-mirror reactor is roughly 10, or about half the minimum Q required of a practical tokamak.

Investigations into the engineering practicality of fusion reactors began in earnest about 10 years ago. There were at least two motivations. One was to determine the dimensions of the more difficult engineering problems. The results often served as a guide to specific further research. The second was to make it clear that certain aspects of the physics of a reactor may compromise the engineering of a particular scheme of confinement. The results often encouraged ideas in physics that helped to mitigate the engineering problems.

The main engineering features of magnetic fusion reactors are common to all confinement schemes. In the first place, most concepts for magnetic fusion reactors call for superconducting magnets to minimize the power required to maintain the magnetic field. Typically a field strength of 50,000 gauss is needed over a volume of from 3,000 to 10,000 cubic meters. (The average value of the earth's magnetic field is about half a gauss.) If nonsuperconducting magnets were used, the power consumed by the magnets would exceed the reactor's output.

Superconductivity is a property that is shared by many metals and alloys. Basically a phase transition takes place when the material is cooled below some critical temperature, so that the electric resistivity of the material drops to zero. In practical alloys such as niobium-titanium in a ratio of one to one or niobium-tin in a ratio of three to one, superconductivity occurs at temperatures less than 20 degrees Kelvin (degrees C.



TOKAMAK scheme of plasma confinement is the one employed in STARFIRE. It requires three groups of magnets. Toroidal-field magnets set up a field whose lines of force are concentric with the torus formed by the plasma. Ohmic-heating magnets act as the primary turns of a transformer; they induce an electric current in the secondary turn, which is the plasma itself. The current heats the plasma. In addition it sets up a poloidal magnetic field: a field whose lines of force are loops at right angles to the toroidal field. Poloidal-field magnets contribute some of the transformer action. In addition they set up a magnetic field whose vertical lines of force counteract the plasma's tendency to spring outward from a torus. The net magnetic field consists of helical lines of force. Each charged particle in the plasma gyrates around a line and advances along it, but since each line is closed (it meets itself after a few turns around the torus) the particle is trapped.

above absolute zero). The alloy is made into filaments, which are twisted together in groups and embedded in a copper matrix. The resulting conductor is then set in additional copper in order to stabilize it and to ensure that the magnet will continue working even if some malfunction causes the loss of superconductivity. A typical current density is from 20,000 to 40,000 amperes per square centimeter. Layers of conductor are generally interleaved with stainless steel supports; layers of insulation intervene between the two. The interleaving is surrounded by a stainless steel dewar, or cooling vessel.

The coolant is liquid helium, which boils at 4.2 degrees K. A common approach is to provide passages for the helium so that the conductor makes contact with a helium bath. Heat generated in the conductor causes the helium to boil, and the resulting vapor carries the heat away. The vapor is then condensed in a refrigerator. A modern liquid-helium refrigerator requires about 500 watts of power to remove one watt of heat deposited at 4.2 degrees K. Nevertheless, the consumption of power by the magnet refrigerators in a reactor generating a million kilowatts of electricity is expected to be modest, on the order of 10,000 kilowatts.

A substantial part of the magnet-design problem is structural. Superconducting magnet coils can be regarded as large pressure vessels containing the energy stored in the magnetic field. In tokamaks and tandem mirrors the amount of energy is from 50 to 100 billion joules. The forces arising from that storage are both in the plane of the coil and across it. In a tokamak, for example, the toroidal component of the magnetic field (that is, the component concentric with the torus) is produced by circular or *D*-shaped coils through which the toroidal plasma passes. The in-plane load produces a radial force on each coil that amounts to about 15,000 pounds per square inch; it places the coil in tension. Meanwhile an out-of-plane load that varies with position acts to overturn the coil.

There is every indication that the investment made so far to develop stable, steady-state superconducting magnets will make such magnets available when fusion reactors need them. Magnets generating fields that change over time at rates as high as 10,000 gauss per second present more difficult problems. The changing fields entail time-varying forces and also induce eddy currents in the coil casings. The eddy currents dissipate as heat in the helium coolant and add to the refrigeration load. Will reactors need time-varying fields? Tandem-mirror reactors will not, but tokamaks might, unless radio-frequency waves prove capable of driving the plasma current up to its operating value. Here, then, is an instance in which the outcome of physics experiments will profoundly influence the technology (perhaps even the engineering feasibility) of a confinement scheme.

A second engineering feature common to all confinement schemes is that auxiliary heating is required to raise the temperature of the plasma. If the plasma carries a current, the resistive heating will help but will generally be insufficient. The two commonest auxiliary-heating techniques employ radio-frequency waves or intense beams of high-energy neutral atoms or molecules.

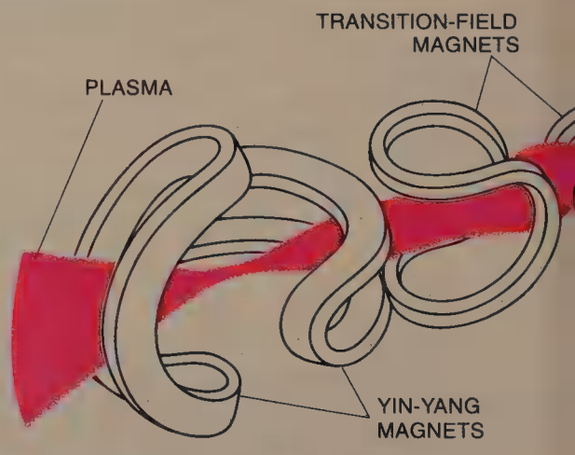
Radio-frequency heating occurs when electromagnetic or electrostatic R.F. waves are converted into thermal energy by a resonant interaction between the waves and the plasma particles. Three frequency regimes are receiving much attention. The first ranges from 50 to 100 megahertz, where the wave frequency matches the ion cyclotron frequency or one of its harmonics. The second ranges from one to three gigahertz, where the wave frequency matches the frequency of fluctuations in the density of ions in the plasma. The third ranges from 50 to 100 gigahertz, where the wave frequency matches the electron cyclotron frequency.

All R.F. heating systems are in essence transmitting stations. They include an R.F. source, an amplifier, a transmission line and a launching structure at or inside the plasma-chamber wall. The source and the amplifier are tubes, klystrons or masers, depending on the frequency regime; the transmission line is a coaxial cable or a waveguide; the launcher is an antenna or a waveguide. The launcher must be hardened against damage from the neutrons and other forms of radiation emitted by the plasma. In addition a dielectric window must be interposed in the transmission line to pass the waves while preventing tritium, which is radioactive, from streaming into the system.

Intense beams of high-energy neutral particles (atoms or molecules) are the alternative method of plasma heating. In some schemes, such as tandem-mirror reactors, they are essential. The neutral particles move unaffected through the reactor's magnetic field and are ionized and trapped by collisions with the plasma's ions and electrons. The plasma is then heated as the fast ions are slowed by further collisions. The beam itself begins as a plasma produced by an R.F. plasma generator or by a high-current discharge. It is typically ionized hydrogen and deuterium atoms, with a significant admixture of diatomic and triatomic ionized molecules. The plasma is accelerated in one or more stages by a series of metal grids that must be continuously cooled and must not distort the beam. The objective is to produce particles moving parallel to one another at a single velocity.

Next the particles are neutralized by passing the beam through a chamber containing cold neutral hydrogen or deuterium gas. The ions each pick up an electron from a cold neutral atom in what are called charge-exchange reactions. The cold neutral atoms, which have now become ions, drift to a wall of the chamber and are collected. Meanwhile the beam passes into a drift tube. The small fraction of ions remaining in the beam is deflected by a magnet. The main beam, consisting of neutral atoms, passes across the reactor's magnetic field and into the plasma.

The technology of neutral-beam systems is made difficult by the combination of energy, power and current that fusion reactors will require. For adequate penetration of the beam into the fusion plasma the beam's energy will have to be 100 to 200 kilovolts. And since the heating of the plasma may require as much as 100 megawatts a beam of, say, 200-kilovolt particles will have to transport the particles at a rate equivalent to a current of 500 amperes and at a current density equivalent to about one ampere per square centimeter. Such needs are enormous in comparison with the attributes of ordinary particle accelerators in the laboratory, whose current at 200 kilovolts is measured in thousandths of an ampere.



TANDEM MIRROR scheme of plasma confinement is an alternative to the tokamak. It relies in part on magnetic fields, in part on the

Nevertheless, the progress in neutral-beam engineering has been truly remarkable. Neutral-beam systems developed in the early 1970's at the Oak Ridge National Laboratory and the Lawrence Berkeley Laboratory delivered about 100 kilowatts of power in a beam with an energy of about 10 kilovolts. The pulse length was very short: less than .05 second. Today beam systems on test stands deliver 10,000 kilowatts of power at an energy greater than 100 kilovolts for periods approaching half a minute.

The heat deposited in the plasma by the auxiliary heating system must eventually come back out, accompanied by the heat that alpha particles have transferred to the plasma. The rate can be predicted: in a reactor intended to generate perhaps 1.2 million kilowatts of electricity it will be between .5 and one million kilowatts. Clearly the engineering of the components surrounding the plasma to handle heat is a major concern.

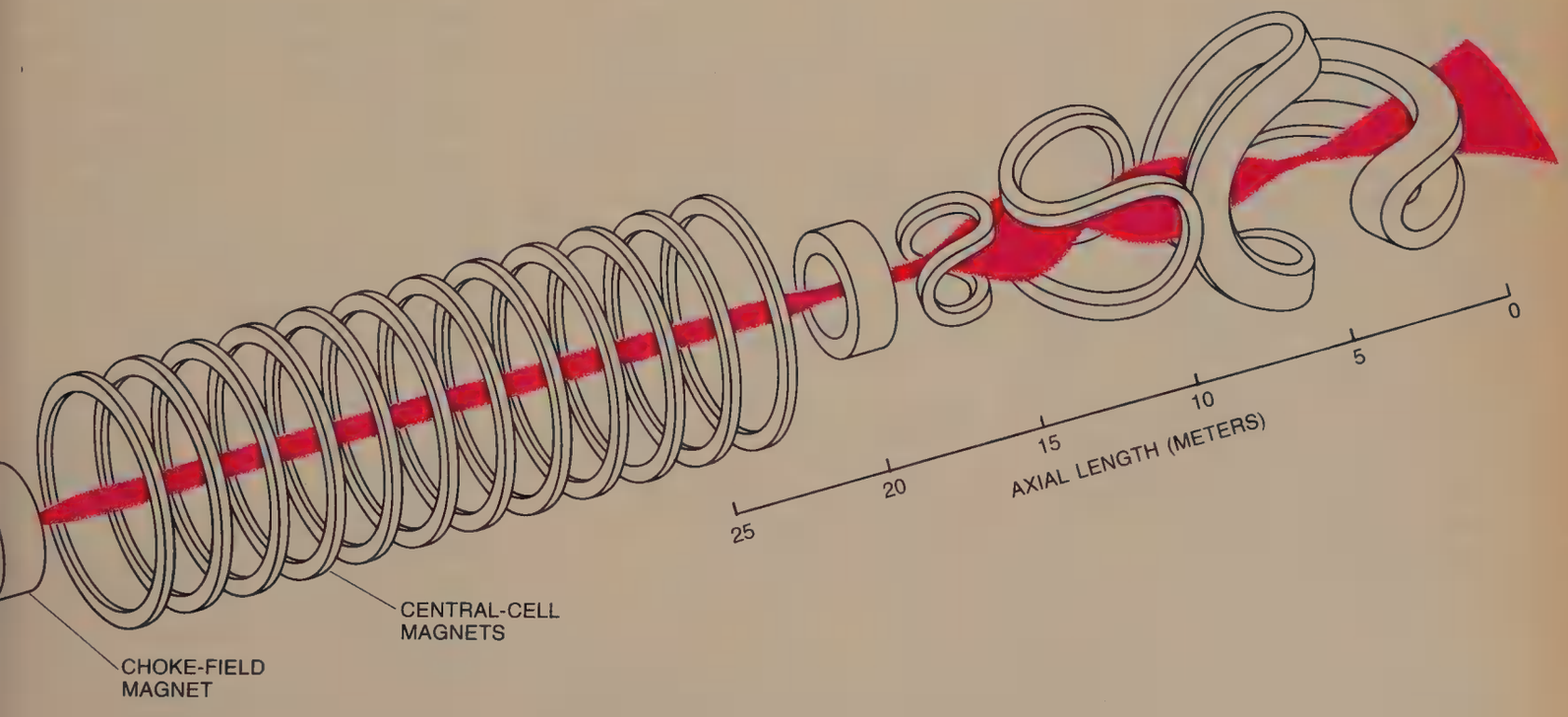
The heat losses will take several forms. For example, electromagnetic radiation (primarily X rays) will be emitted by the plasma. Impurity atoms in the plasma can greatly increase the rate of such radiation; hence attention is being given to maintaining the purity of the plasma and minimizing the erosion of the first wall (the inner surface of the plasma vessel) and the various in-vessel structures. In open-ended designs such as tandem mirrors charged particles are lost primarily from the ends of the plasma; they can be handled at a place remote from the plasma vessel.

In toroidal reactors the situation is different: charged particles diffusing radially outward from the center of the plasma represent the major loss. Such particles will interact with a limiter, a specially designed device capable of handling a heat loading as great as 10 million watts per square meter. Situated inside the first wall, it is a material contact point for the plasma. In some designs a special magnetic-field configuration will divert plasma particles, and the heat they carry, from the plasma vessel to a "divertor region," where the heat can be removed from the reactor by collector plates and the backstreaming of neutral gas and impurities can be minimized.

The related problem of first-wall erosion under the impact of plasma particles has been controlled in recent experiments by making the limiter out of materials such as graphite (carbon) that have a low atomic number. (The emission of radiation by impurities in a plasma tends to increase with increasing atomic number.) Chemical reactions between hydrogen and carbon, however, are likely to prevent the use of graphite inside the first wall of a reactor. Workers at the Sandia National Laboratory and elsewhere are developing special materials such as silicon carbide and titanium carbide for use as coating on components that must withstand both a heat flux and a plasma-ion bombardment.

I come now to the production of power by the reactor. Only a fifth of the energy released in the reaction of a deuteron and a triton is taken away by an alpha particle. The remaining four-fifths is taken away by a 14-million-electron-volt (14 MeV) neutron, which flies freely out of the plasma, into the first wall and the surrounding region: the blanket. The blanket has two functions. First, it recovers the neutron's energy. A 14-MeV neutron typically travels about 10 centimeters between collisions with the nuclei of solids or liquids. At each collision the neutron transfers energy to the target atom, its trajectory is changed and heat is deposited locally in the blanket. Coolant flowing through the blanket removes the heat and transports it to a relatively conventional power plant consisting of heat exchangers, steam generators and turbines. The neutron can also be absorbed by a nucleus, transforming the atom into one that may be stable or radioactive. The neutron flux attenuates with distance from the first wall; hence the blanket is typically 60 to 80 centimeters thick. Neutrons leaking from the back of the blanket are captured by a shield designed primarily to protect the superconducting magnets from radiation damage and excessive heating.

The second function of the blanket is



electrostatic field arising from the distribution of electric charge in the plasma. Here the required magnets are shown. Ring-shaped magnets surround a central cell of plasma. Then at each side of the cell is a choke-field magnet, a pair of transition-field magnets and a pair of magnets in what is known as a yin-yang configuration. The scale toward the right matches the scale in the chart on the following page.

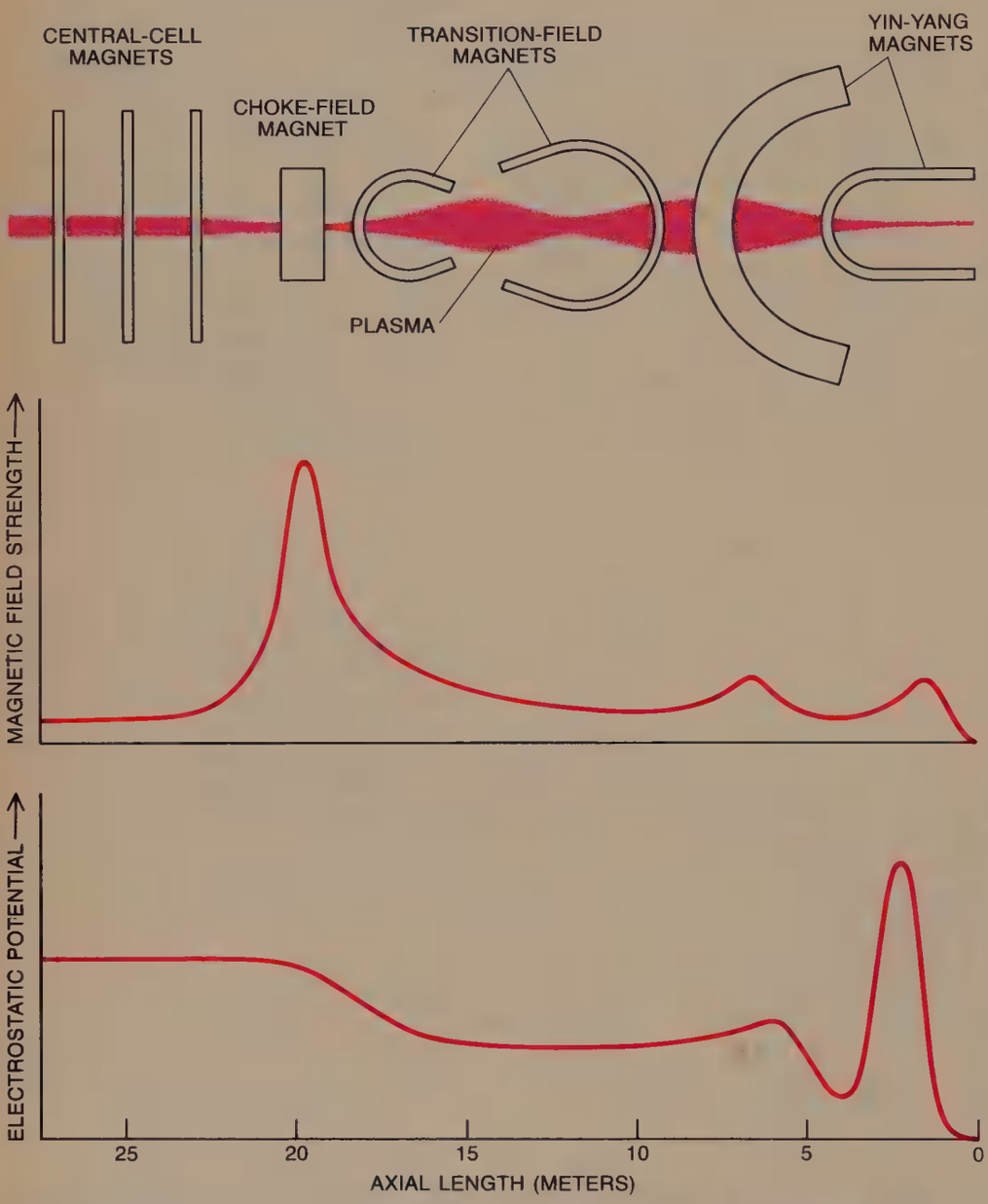
to make fuel for the reactor. Deuterium is plentiful (it is found at a ratio of about one part in 5,000 in ordinary water), but tritium is radioactive, with a half-life of 12.35 years, and so it is not abundant in nature. (Minute amounts of it are produced by cosmic-ray reactions in the upper atmosphere.) Fortunately tritium can be "bred." Specifically, the 14-MeV neutron produced in the fusion reaction can be used to induce nuclear reactions with the two naturally occurring isotopes of lithium: lithium 6 and lithium 7. Neutrons of any energy can be captured by lithium 6 to produce tritium and helium; neutrons with an energy greater than 2.87 MeV can be captured by lithium 7 to produce tritium, helium and another, slower neutron. The breeding ratio of the reactor (the number of tritons produced per triton consumed) depends on details of the design, but it can be as high as 1.5.

The ingredients of a blanket design are thus the structure, the coolant and the tritium-breeding material. Leading candidates for the structure are stainless steels, ferritic or high-manganese steels and vanadium alloys. Leading candidates for the coolant are water, a gas (such as helium) or a liquid metal. The leading liquid metals are lithium and lithium-lead compounds and eutectics. Liquid lithium is attractive because it has a low melting point (186 degrees C.), excellent heat-transfer properties, good chemical compatibility with most structural alloys and acts simultaneously as the tritium-breeding material.

The tritium could be recovered from a side stream of the lithium. Since it would be in the form of lithium tritide, one recovery method would be to bring the side stream into contact with an equal volume of a molten salt such as a mixture of lithium fluoride, lithium chloride and lithium bromide. The lithium tritide would tend to pass into the salt, but the two are readily separated because the lithium tritide is three to four times lighter. The tritium is then recovered by electrolysis. The overall process is a simplified version of the molten-salt reprocessing technique that was developed to reprocess fission-reactor fuel.

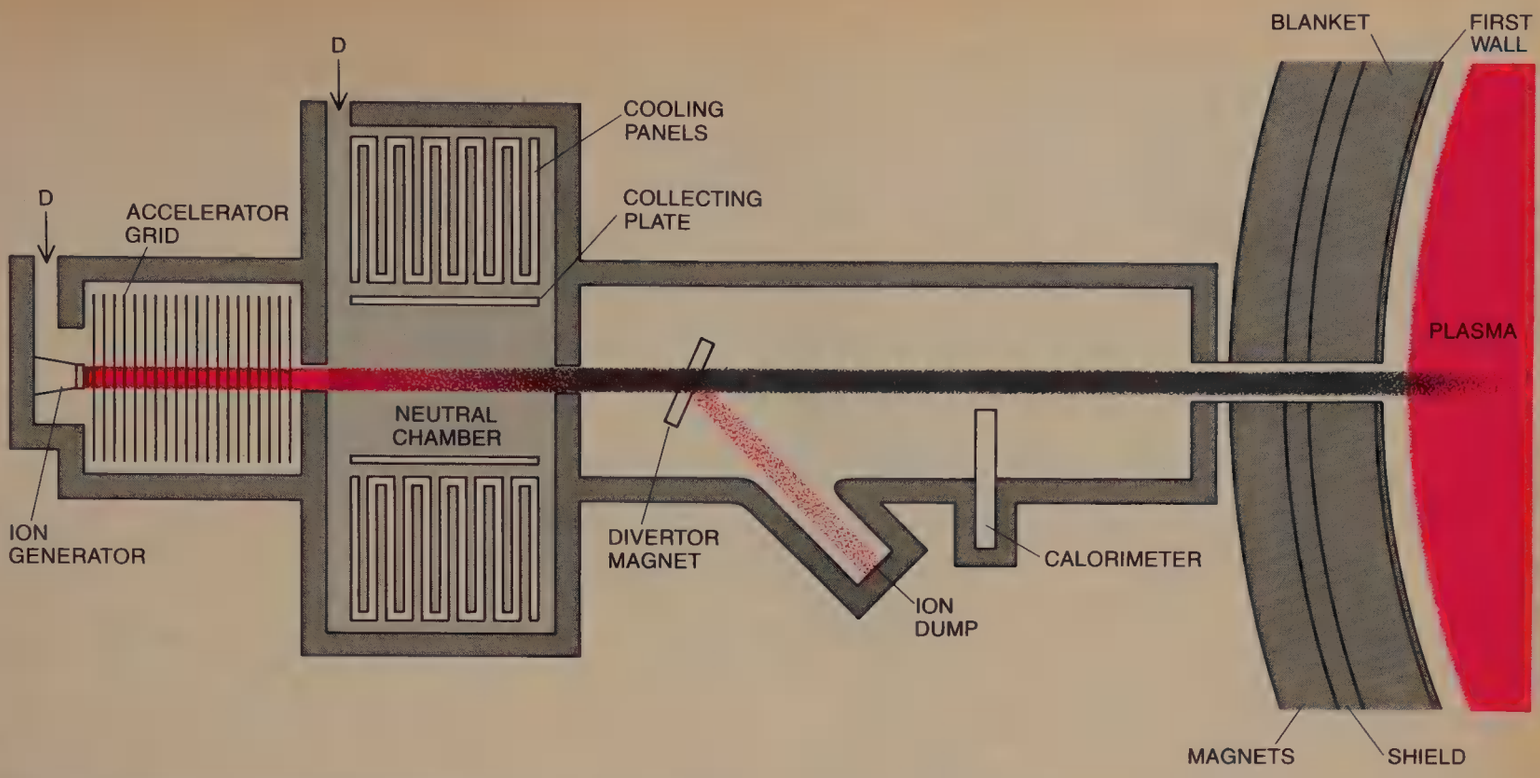
Ceramic compounds including lithium oxide, lithium aluminate and lithium silicate are alternative tritium-breeding materials. Two of their advantages are that the amount of lithium in the reactor can be modest and that the compound is chemically stable. On the other hand, a breeding ratio greater than one will be difficult to achieve. In lithium aluminate ($LiAlO_2$), for example, aluminum and oxygen atoms compete with the lithium to scatter neutrons or absorb them. What is needed is a neutron multiplier. Positioned in the blanket just behind the first wall, it converts one very fast neutron into two slower neutrons by means of a nuclear reaction. The best neutron multipliers are beryllium and lead.

The recovery of tritium from a ceramic is difficult since it relies on the diffusion of tritium out of the solid. The typical diffusion rate is low, and so the ceramic must be fabricated in pellets composed of grains whose size is on the order of 10^{-3} millimeter. Yet the ceramic must not sinter, even when its temperature approaches or perhaps exceeds 1,000 degrees C. The tritium diffusing from the ceramic would be transported from the blanket in a purge stream of helium gas flowing through the pellet beds. Whether the breeder will perform according to the design is a major question likely to be answered only through experimentation.



FIELD STRENGTH AND ELECTRIC POTENTIAL at each end of a tandem-mirror reactor result from magnets, from the heating of the plasma by beams of high-energy electrically neutral particles and from the pumping of ions out of certain regions of the plasma. The charts show the strength of the magnetic field and the electrostatic potential toward either end of a tandem mirror. Ionized atoms in the plasma tend to be trapped in the central cell by the magnetic peak at the choke magnet. In essence they are reflected by a bottleneck in the magnetic field. The bottleneck acts as a mirror. The ions that do get through the mirror tend to bounce off the electrostatic peak inside the yin-yang configuration. Electrons in the plasma tend to be repelled as they approach either side of the electrostatic valley inside the transitional region. The electrons isolated electrostatically in the end cell can thus be heated efficiently.

Certain engineering problems affect several components of a reactor. One such problem is neutron radiation damage. When a neutron is scattered or absorbed in a collision, some of its energy is transferred to the atom that it struck. The atom recoils from the site of the collision and displaces other atoms, some of them permanently. These other atoms, which are called interstitials, leave vacancies behind. The interstitials can form loops, and so can the vacancies. Indeed, the vacancies can form permanent voids, particularly if some agent such as gas is present to stabilize them. In the metallic components of a fusion reactor gas is indeed produced. For example, hydrogen and helium are produced when neutrons with an energy



NEUTRAL-BEAM INJECTION is one way to heat a fusion plasma. The particles in the beam (often deuterium atoms) are ionized and accelerated. Then they capture electrons from cold deuterium atoms, becoming neutral again. They can now cross a magnetic field, enter the plasma and transfer energy by collisions. An alternative is to heat the plasma with electromagnetic radiation at radio frequencies.

greater than about 5 MeV interact with metal atoms. In the course of the interactions some of the metal atoms can be transformed into radioactive isotopes of the same elements or different ones. As a result of all these changes the material's properties change. Commonly the material hardens and loses ductility.

The flux of 14-MeV neutrons through the first wall of a fusion reactor is measured in units of megawatts per square meter. A flux of 4×10^{17} neutrons per square meter per second represents a wall loading of one megawatt per square meter. The actual wall loading may be from three to five times as great. It is possible to calculate how many times per year an atom in the wall will be displaced from its site in the lattice of a material for each megawatt-per-square-meter load. The number is staggering. In a stainless steel first wall every atom will be displaced 10 to 20 times. Since practically all the displaced atoms come to rest at some vacant site, the steel does not turn to jelly. The problem is nonetheless clear. The same load will also produce about 150 helium atoms and 500 hydrogen atoms per million metal atoms per year.

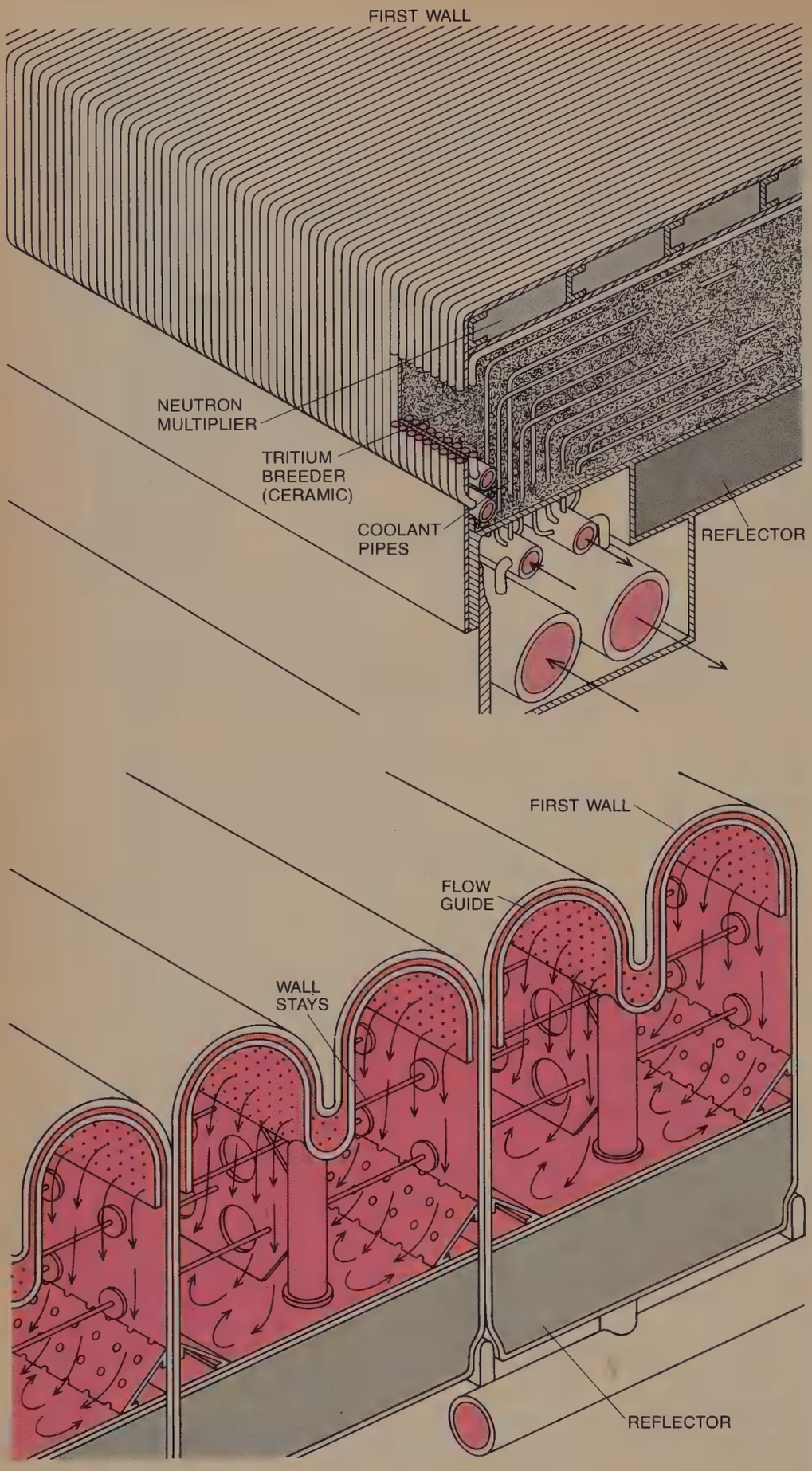
After plasma confinement the changes in materials due to radiation damage are perhaps the most challenging problem confronting the engineers of fusion reactors. Research has begun with the irradiation of metal samples in fission reactors such as the High Flux Isotope Reactor at Oak Ridge, which is designed to produce great numbers of neutrons having a wide spectrum of energies. When the sample is an alloy including nickel, thousands of helium atoms per million metal atoms are produced, and changes such as swelling and embrittlement have been observed. A preliminary estimate is that the first wall and the blanket in a working fusion reactor would retain adequate properties for from three to six years. The implication is that fusion reactors must be designed with a first wall and blanket that can be replaced periodically throughout the life of the plant. Clearly much further research is needed. A special high-flux 14-MeV-neutron-source facility, the Fusion Materials Irradiation Test, is being considered for construction by the U.S., possibly in collaboration with Western European countries and Japan. It will be an important part of a program to develop adequate materials for fusion reactors.

In both tokamaks and tandem mirrors the rate of leakage of thermonuclear fuel from the plasma is 10 to 20 times greater than the rate of fuel consumption. This makes fuel recovery and recycling an imperative. The recovery is a challenge because the pumping rate must be high, ranging from one to 10 million liters per second, and the vacuum left behind in the reactor vessel must be impressive, about 10^{-7} atmosphere. Currently cryosorption pumps operating at the temperature of liquid helium appear to be the best of the technical alternatives.

Cryopumps include an array of molecular sieves that are cooled by liquid helium to a temperature of 4 degrees K. The sieves are surrounded by chevrons that are cooled by liquid nitrogen to a temperature of 77 degrees. Volatile gases such as hydrogen, deuterium and tritium condense on the sieves; less volatile gases such as oxygen, nitrogen and methane are adsorbed by the chevrons. Helium can also be captured if liquid argon is sprayed onto the surfaces of the pump; the argon traps the helium by overlaying it. Cryopumps have a high pumping speed but require periodic removal of the gases they trap. Moreover, the trapping entails a high inventory of tritium in the reactor. In response to these difficulties designs have been developed in which devices such as magnetic divertors or pump limiters raise the gas pressure at the beginning of the pumping ducts to about 10^{-5} atmosphere. Pumps that do not trap fuel can then be employed.

Once the exhaust stream is purged of its elemental impurities (including the helium "ash" produced by the fusion reactions) it is necessary to separate the unburned deuterium and tritium so that they can be combined in proper proportion with tritium recovered from the blanket. The separation is most straightforwardly accomplished by distillation at cryogenic temperatures, making use of the difference in the boiling points of H_2, DH, D_2, HT, DT and T_2. The least volatile is H_2, with a boiling point of 20.39 degrees K.; the most volatile is T_2, with a boiling point of 25.04 degrees. Acceptable purity can be obtained using distillation columns with about 60 stages.

The reconstituted deuterium-tritium mixture must then be reinjected into the



FIRST WALL AND BLANKET vary in design depending on the choice of coolant and of tritium-breeding material. In the design at the top the coolant is water (*color*) and the breeder is lithium aluminate or lithium oxide, each a solid ceramic. Neutrons escaping the fusion plasma transform lithium in the ceramic into tritium, helium and additional neutrons. The tritium then diffuses out of the ceramic and is flushed out of the blanket by a stream of helium gas. The production of tritium is increased by a neutron reflector (graphite) under the blanket and by a neutron multiplier (beryllium or lead) above it. Nuclear reactions in the multiplier absorb fast neutrons from the plasma and emit twice as many slower ones. In the design at the bottom the coolant and the breeder are the same: they are liquid lithium (*color*). The tritium is recovered by removing it from the coolant on its way to the reactor's heat exchangers. Both designs in the illustration were prepared as part of the STARFIRE study of the design of a tokamak.

plasma. In current experiments with hydrogen plasmas puffs of gas are aimed at the edge of the plasma, which readily ingests them. In a reactor the ingestion may not be as effective; the gas atoms would be ionized no farther than 10 centimeters in from the plasma's edge. Ten or more puffing systems would be required, each consisting of a nozzle and a fast-acting valve. Each would produce puffs lasting 50 to 100 milliseconds.

The alternative is the high-speed injection of frozen fuel pellets. The process has been demonstrated in an experimental tokamak at Oak Ridge and seems sure to work in a commercial reactor. Each pellet is truly a snowball in hell: made at a temperature of less than the freezing point of deuterium (18.73 degrees K.) and tritium (20.62 degrees), it would be shot into a plasma whose temperature is tens of millions of degrees. It is envisioned that each pellet would be two to four millimeters in diameter and would contain 10 to 15 percent of the total number of fuel atoms in the plasma at any instant. The pellets would be accelerated to a speed of about two kilometers per second and then injected at a rate of 10 to 20 pellets per second. Except for the injection rate these operating parameters have already been achieved.

Two types of injectors have been proposed. Centrifugal injectors consist of a high-speed rotating plate on which two *U*-shaped tubes are mounted. A feed line cooled by liquid helium extrudes solid fuel into one of the openings of each *U*, and a knifing action cuts the fuel to form a pellet inside the tube. The rotation of the plate then accelerates the pellet so that it emerges from the tube at high velocity. In experiments at the Max Planck Institute for Plasma Physics in Garching and at Oak Ridge one-millimeter deuterium pellets have been accelerated to speeds of 290 meters per second. The devices should ultimately be able to accelerate pellets to speeds of from two to five kilometers per second. A high repetition rate is a built-in characteristic of the scheme.

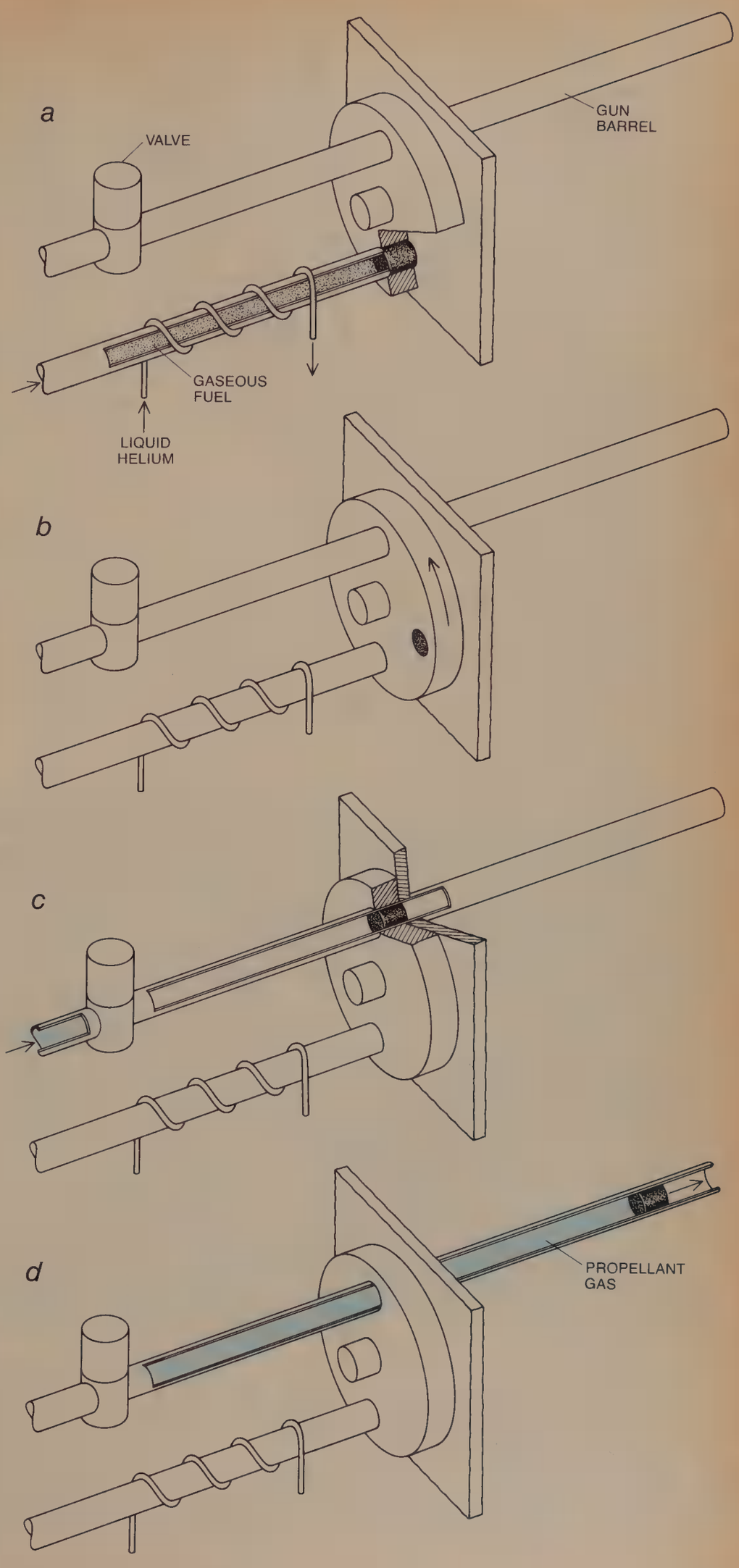
Gas guns are also being developed. Indeed, they have accelerated one-millimeter pellets to a speed of one kilometer per second. In a design worked out at Oak Ridge a disk cooled by liquid helium is in a copper housing; one-millimeter holes have been drilled through the disk. Gaseous fuel freezes in each hole, and the disk is rotated so that the resulting slug of fuel comes into line with a fast-acting valve at the end of a gun barrel 16 centimeters long. The valve releases helium gas, which has been compressed to a pressure of 30 atmospheres. The gas propels the slug. Performance can be improved by using hydrogen as the propellant and by increasing the length of the barrel.

Since tritium is radioactive one must

ensure that throughout the fuel cycle the tritium is well contained in the reactor. The required degree of containment can be determined by a simple calculation. The inventory of tritium in a power-producing reactor is expected to be about 10 kilograms, or about 100 million curies of radioactivity. (One curie is equivalent to 30 billion nuclear disintegrations per second.) A safe and allowable escape rate has not yet been established, but research groups contemplating reactors have used figures ranging from one to 10 curies per day. This means that the reactor must routinely lose no more of its tritium per day than one part in 100 million.

The containment can be achieved by providing at least three physical barriers at each point in the tritium handling. The first barrier is the structural element containing the tritium, say a pipe. Then comes a barrier surrounding the structural element. For example, a pipe could be within a second pipe or a group of tritium-handling components could be within a glove box. The third level of containment would be the reactor hall, the rooms containing the tritium-processing equipment and so on. Each would contain a tritium-cleanup system. It is estimated that a kilogram of tritium that had escaped in the reactor building could be removed in less than a day. The Tritium System Test Assembly, an experimental mockup of the entire tritium fuel cycle, has been constructed at the Los Alamos National Laboratory and is scheduled to begin operation soon. It should provide the information necessary to ensure the safe handling of tritium in future experimental facilities and then in reactors.

There is only one other source of radioactivity in a fusion reactor: the reactor itself, and in particular the first wall, the blanket and the shield, are made radioactive by the flux of neutrons through them. The amount and the nature of the radioactivity will depend on the materials of which the reactor is made. A useful figure of merit is the biological hazard potential, or BHP. The Federal Government lists the maximum permissible concentration of various radioisotopes in air or in water; the BHP is the ratio of the actual radioac-



FUEL INJECTION can be accomplished by propelling frozen pellets of deuterium and tritium into the plasma. In an injector devised at Oak Ridge National Laboratory a stream of gaseous fuel is frozen by liquid helium so that a slug of solid fuel is emplaced in a hole in a disk (*a*). The disk rotates (*b*), bringing the slug into line with a valve at the end of a gun barrel (*c*). The valve releases spurts of propellant (*color*): high-pressure helium gas (*d*). The injector in a fusion reactor will have to give pellets a velocity of two kilometers per second.

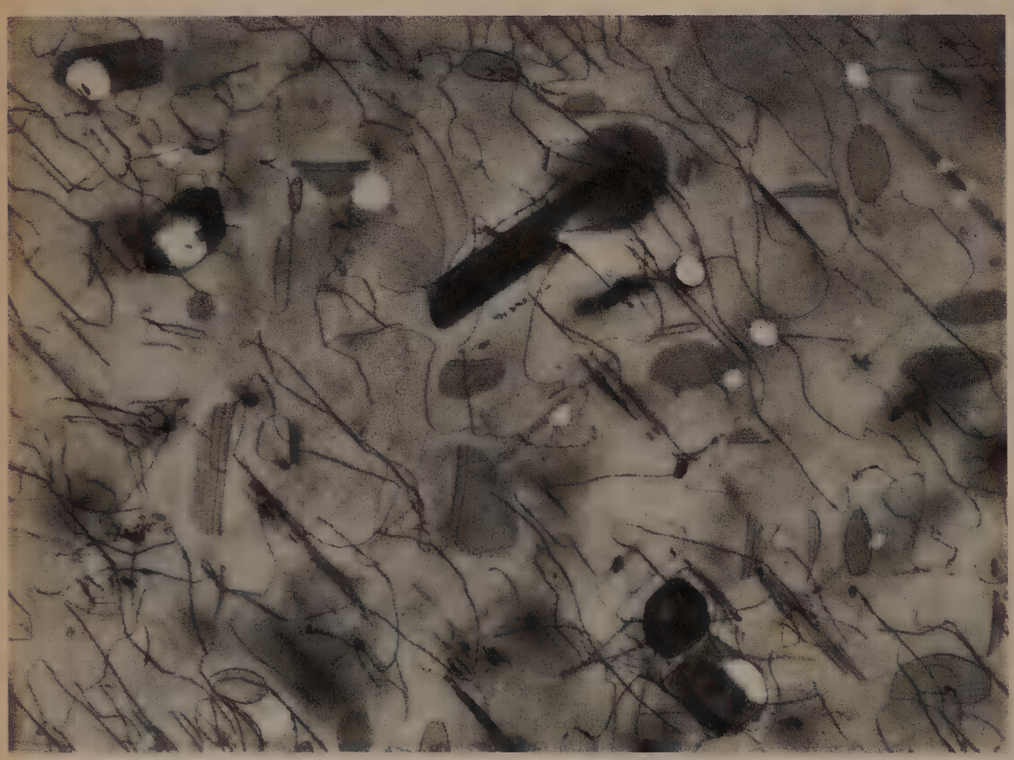
tivity (in curies per unit volume) to the Government's permissible level. For all materials of practical interest the BHP for a fusion reactor will be at least a factor of 10 below the BHP for a fission reactor. In many cases the factor of improvement will be more than 100.

Perhaps it is more important to ask whether the waste materials from a fusion reactor could be disposed of by shallow-land burial. Federal guidelines for such burial have recently been proposed. Under them three classes of waste are described. Class *A* waste is material that attains an acceptably safe level of radioactivity after no more than 10 years of shallow burial. "Acceptably safe" is defined to mean that a person inadvertently entering the burial site and staying there would receive a yearly dose of radiation no greater than five times the natural background level. Class *B* waste is chemically stabilized material that attains an acceptably safe level within 100 years. The material must stay stable and must be buried so that the dose to a person at the site would be no more than a few percent above the background level. Class *C* waste is chemically stabilized material that attains an acceptably safe level within 500 years. The material must be buried under at least five meters of earth and the site must have natural or engineered barriers to trespass. Materials that fail to meet these prospective guidelines would presumably require a more extreme disposal method such as deep storage in a suitable geologic formation.

Fusion reactors can be built of materials that would be suitable for shallow-land burial. Specifically, calculations show that fusion reactors built of an iron-manganese-chromium steel will yield waste materials of Class *C*. And if the reactors could be built of titanium, vanadium or aluminum alloys or of ceramics such as graphite and silicon carbide, the waste materials would be of Class *B*. Moreover, it may be possible to modify alloys that have useful mechanical properties but a tendency toward high radioactivity into ones that at least meet the Class *C* criteria. Attempts to eliminate nickel and molybdenum from steel alloys are an example of such an effort.

In the longer run it may even be possible to tailor the isotopic composition of a material so that its mechanical and chemical properties are unchanged while its tendency toward radioactivity is altered dramatically. Again consider molybdenum. In most steels it imparts high-temperature strength. It is typically 1 percent of the alloy. The natural isotopes of molybdenum that react with neutrons to form long-lived radioactive elements in a fusion reactor are those with atomic weights of 92, 94, 98 and 100. Suppose the molybdenum for the steel in a reactor consisted of only the remaining natural isotopes: 95, 96 and 97. Then the molybdenum could be added and the steel would nonetheless meet the criteria for Class *C* waste.

The economics of isotope separation are not yet firmly established; the present techniques were developed to separate only isotopes of uranium. A new technique being developed at TRW, Inc., could be used to separate lighter elements such as nickel and molybdenum. Moreover, a goal of $220 per kilogram has been established for uranium enriched to 3 percent in the isotope U-235. Suppose the isotopic tailoring of molybdenum proves to cost four times that much. The cost of a fusion-reactor blanket made with tailored-molybdenum steel instead of an untailored alloy would be increased by less than 5 percent.



NEUTRON-RADIATION DAMAGE that the materials in a fusion reactor will suffer was simulated by placing a sample of stainless steel in the core of a research fission reactor at a temperature of about 590 degrees Celsius and exposing it for a year to a high flux of neutrons. The total flux (1.9×10^{26} neutrons per square meter) displaced each atom in the lattice of the steel an average of 10 times. Some of the vacated sites in the lattice aggregated into voids (*white areas*); some of the displaced atoms aggregated into regions called dislocation loops (*medium gray areas*). In addition the chemical elements in the steel segregated (*dark gray areas*). As a result the steel swelled and became less ductile. The electron micrograph was made by F. W. Wiffen, J. O. Stiegler and E. E. Bloom at Oak Ridge; it enlarges the steel 110,000 diameters.

Two comprehensive design studies, one of a tokamak, one of a tandem mirror, have recently been completed. Each provides a glimpse of what fusion reactors may be like. The tokamak study, STARFIRE, was led by the Argonne National Laboratory in association with the McDonnell-Douglas Astronautics Company, GA Technologies and the Ralph M. Parsons Company. It took two years and cost more than $2 million, and it contemplated the design of a 1,200-megawatt fusion reactor. The device is impressively large: its toroidal plasma has a major diameter of 14 meters and a cross section of 4.2 meters. The plasma is confined by superconducting magnets. The toroidal field is generated by 12 such magnets that form hoops around the plasma's cross section. They generate a maximum field strength of 111,000 gauss, and together they store 50 billion joules of magnetic-field energy.

The plasma in STARFIRE is fueled by a combination of gas puffs and pellets, is heated by R.F. waves and carries a current of 10 million amperes. The current is sustained by R.F. waves in a process that has not yet been fully demonstrated. Without it the plasma burn would have to be allowed to terminate periodically, perhaps every 20 minutes or oftener, while the magnets representing the primary turns of a transformer were reset. In the course of the resetting the reactor chamber and the blanket would cool somewhat, thereby changing the stresses in the materials. The resulting mechanical fatigue would limit the life of the structures. The R.F. driving comes at a price, however; the price is power consumption. It is estimated that 150 megawatts, or more than 10 percent of the electric power produced by the reactor, must be recirculated to supply the R.F. system. The net efficiency of

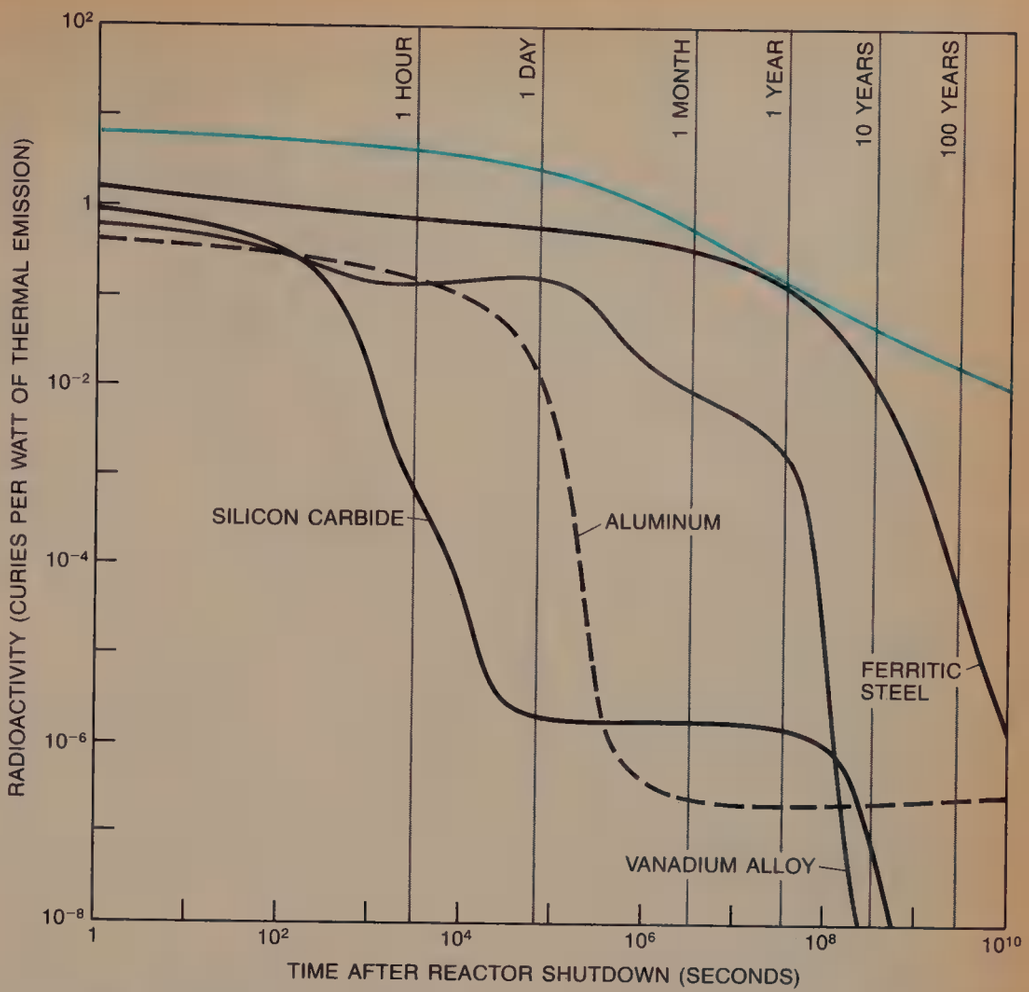
STARFIRE (30 percent) will have to be improved in future tokamak designs.

The blanket in STARFIRE is cooled by water pressurized to 2,200 pounds per square inch. The water enters the reactor at a temperature of 280 degrees C., picks up heat as it circulates through the blanket and leaves at a temperature of 320 degrees. As it circulates it passes through stainless steel tubes surrounded by tritium-breeding ceramic. A layer of zirconium-lead along the inner surface of the blanket serves as a neutron multiplier. Even so, the breeding ratio of the reactor is estimated to exceed one by no more than 5 percent. That is likely to be inadequate. Alternative designs are being developed.

The tandem-mirror study, the Mirror Advanced Reactor Study (MARS), is being led by the Lawrence Livermore National Laboratory in association with TRW, the General Dynamics Corporation and the University of Wisconsin. After two years it is nearing completion. It contemplates a long, linear reactor with a central plasma cell 129 meters long between a pair of end cells acting as mirrors that together add 48 meters. The central cell holds 75 cubic meters of plasma; it yields a continuous power output of 2,575 megawatts. The end cells are sustained by the continuous injection of neutral beams and R.F. heating. The central cell is fueled by gas puffs and pellets. A "halo" plasma surrounding the main cylindrical plasma keeps impurities from entering.

The magnets in MARS include 41 circular superconducting rings in the central cell; they generate a field of 47,000 gauss. In addition superconducting end-cell yin-yang magnets produce a maximum field of 75,000 gauss, and a magnet of special construction, combining water-cooled nonsuperconducting magnets and a superconducting magnet, produces a "choke field" of 240,000 gauss between the central cell and each end cell. A field of such intensity cannot be produced over large volumes by any wholly superconducting magnet known today. The entire magnetic field stores about 50 billion joules. Power is required to drive the nonsuperconducting magnets and refrigerate the superconducting ones, to circulate coolant, to sustain the end-cell plasmas and to operate auxiliary equipment. Indeed, 20 percent of the power produced by the reactor must be recycled. Nevertheless, the net efficiency of the reactor is 35 percent.

Three features of MARS contribute to that efficiency: the electrostatic barrier that separates hot electrons in the end cells from somewhat cooler ones in the central cell, the use of direct convertors to make electricity from the energy of charged particles escaping from the ends of the reactor and the use of a high-temperature coolant, lithium-lead, which enters the blanket at a temperature of 350 degrees C. and leaves at 500 degrees. Not all these features have been demonstrated experimentally. In addition the physics of tandem mirrors is not yet as advanced as the physics of tokamaks. Thus MARS embodies more assumptions than STARFIRE.



DECLINE OF THE RADIOACTIVITY induced by neutrons in the materials of a fusion reactor varies markedly from one material to another (*black curves*), suggesting that materials such as alloys could be tailored in composition so that they retain useful structural properties while becoming less hazardous. In general the decline of radioactivity of a fusion reactor is far more rapid than the decline of the radioactivity inherent in nuclear fission, exemplified by radioactivity in waste from one type of fission reactor, the liquid-metal fast breeder (*colored curve*).

The electrostatic barrier for electrons is an example. It is a region of low electrostatic potential that repels electrons on either side. Positive ions are attracted to the region, and if they are not removed, the dip will disappear. The designers of MARS have therefore incorporated a removal technique called drift-orbit pumping. So far it is entirely theoretical. Special half-moon-shaped coils modulate the magnetic field strength in the end cells by .03 percent at a frequency of about 100 kilohertz. The modulation induces ions to drift radially outward until their orbits extend beyond the realm of electrostatic and magnetic trapping. At that point they escape through the end of the cell.

The designs for tokamaks and tandem mirrors differ in many details. Sometimes the differences result from the need for unique technologies such as direct convertors. The designs have much in common, however, including technologies for vacuum pumping, fuel handling, superconducting magnets, plasma heating and fuel injection. Both STARFIRE and MARS point toward reasonable concepts for fusion reactors, and both show clearly where improvements are required.

Is magnetic fusion practical? The question cannot yet be answered definitively, but sufficient groundwork has now been laid so that the answer can be determined. The U.S., the Western European countries, Japan and the U.S.S.R. all have specific engineering development programs under way, and all of them participated in a multinational study to advance the design of tokamak reactors (INTOR). Perhaps a bilateral or multilateral agreement will facilitate the next major experimental steps. If it does not, the experiments now in progress are sure to be markedly upgraded. Either way, the coming decade will tell much about magnetic fusion engineering.

The Large-Scale Structure of the Universe

Across billions of light-years space is a honeycomb of galactic superclusters and huge voids. The structure may result from perturbations in the density of matter early in the big bang

by Joseph Silk, Alexander S. Szalay and Yakov B. Zel'dovich

Astronomers have long recognized that the distribution of matter on a cosmic scale must somehow bear the imprint of a very early stage in the history of the universe. A consistent account of that distribution and its evolution must be developed within the context of the big-bang theory, since there is almost universal consensus among cosmologists and astrophysicists that the big bang provides an empirical framework within which all cosmological issues can be examined. According to the big-bang theory, the universe began as a singular point of infinite density some 10 to 20 billion years ago and pulsed into being in a vast explosion that continues to this day. In the simplest version of the theory the universe expands everywhere uniformly from the singular point. The uniformity of that expansion accounts remarkably well for much important observational evidence: Extragalactic matter recedes from our galaxy at a rate that varies smoothly with its distance, and a cold bath of radiation in the microwave region of the electromagnetic spectrum pervades the sky at a temperature that varies over a few angular degrees to less than one part in 30,000. In spite of these successes, there is compelling evidence that the expansion is not precisely uniform. If it were, matter would fail to coalesce and the universe would become an increasingly rarefied gas of elementary particles. The stars and the galaxies would not exist.

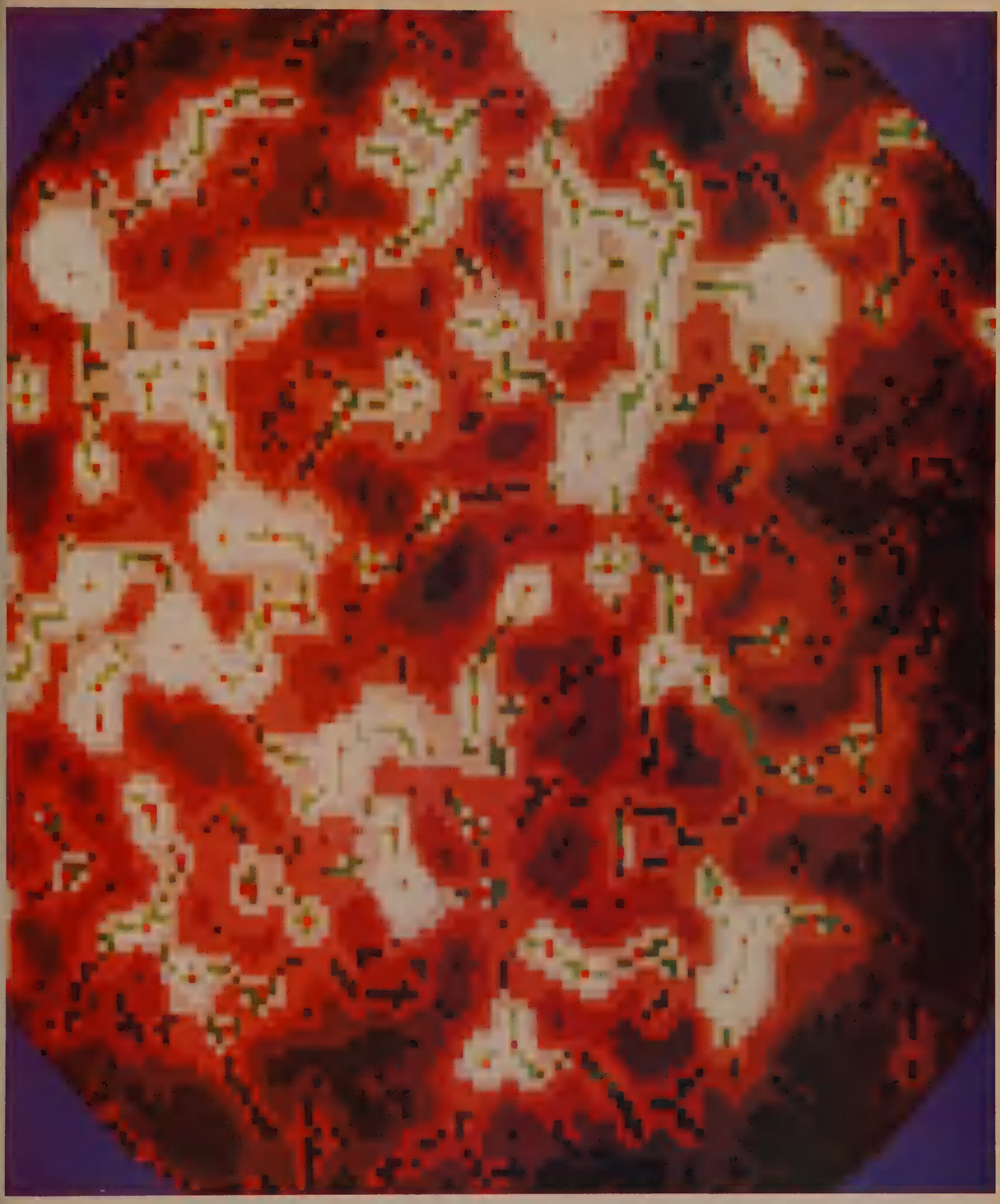
In order to account for structure in the present state of the universe the big-bang cosmologist must therefore acknowledge some degree of clumpiness early on. Such early inhomogeneities might be smooth and nearly indistinguishable against the homogeneous background; small fluctuations in the curvature of the early universe would take the form of slight compressions or rarefactions of matter and energy from region to region throughout space. The amplitude of the fluctuations must be large enough (that is, the variation from the average density must be great enough) to grow into the currently observed aggregations of matter in the time since the universe began; precisely what that amplitude must be, however, is a matter of much theoretical delicacy. If the initial fluctuations were too large, they would cause variations in the temperature of the microwave background radiation that are not observed. Furthermore, the fluctuations must give rise to the structures of relatively special scale that make up the universe and not to structures of arbitrary size. Stars, galaxies, clusters of galaxies and even superclusters, or clusters of clusters, can now be identified, but at scales of mass larger than that of the supercluster the universe is fairly uniform.

The convergence of cosmology and the physics of elementary particles has recently made it possible to satisfy all these requirements without making any strong assumptions about the early state of the universe. In particular, no appeal is made to any special scales or patterns of mass and energy at the outset of the expansion, and no new forces are invoked. What is assumed is that soon after the beginning of the big bang there were small variations in the density of matter and energy everywhere in the universe. The variations were the result of superposing low-amplitude, sinusoidally varying fluctuations in the density at every possible wavelength, or scale of length; the amplitudes of the fluctuations were distributed randomly, and so the resulting variations in density were random and chaotic. Thereafter the present structure of the universe could have evolved according to reasonably well-understood principles of physics.

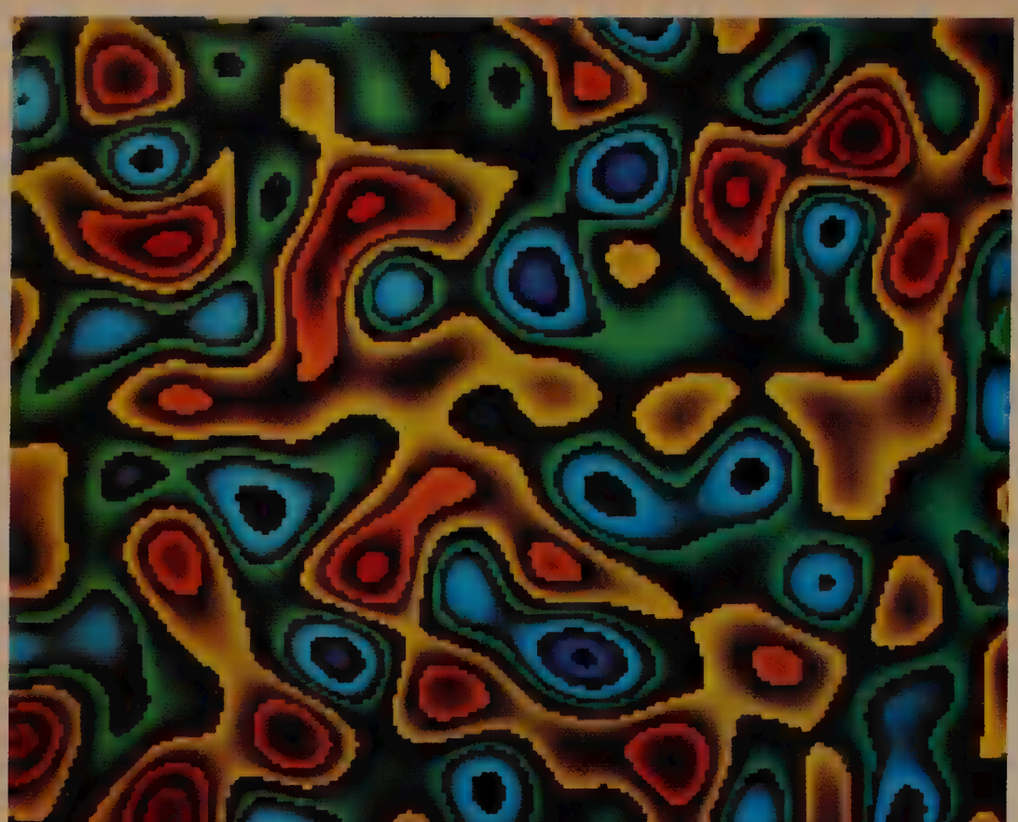
As the universe expanded, the random, free streaming of elementary particles in all directions suppressed all the initial fluctuations below a critical size: the only fluctuations that survived were those that compressed or rarefied masses at least 10^{15} to 10^{16} times the mass of our sun. Gravity then caused some of the compressed masses to contract preferentially along one of the three spatial axes. The initial spectrum of fluctuations thereby gave rise to gigantic, irregular clouds of gas that resembled flattened pancakes. Where the pancakes intersected, long, thin filaments of matter took shape. Some of the clouds remained intact; others broke up to form galaxies and clusters of galaxies. The emergence of an appropriate characteristic scale for the fluctuations was first explained by one of us (Silk). The gravitational formation of the thin layers of matter was discovered by another of us (Zel'dovich). We shall refer to this model as the pancake theory.

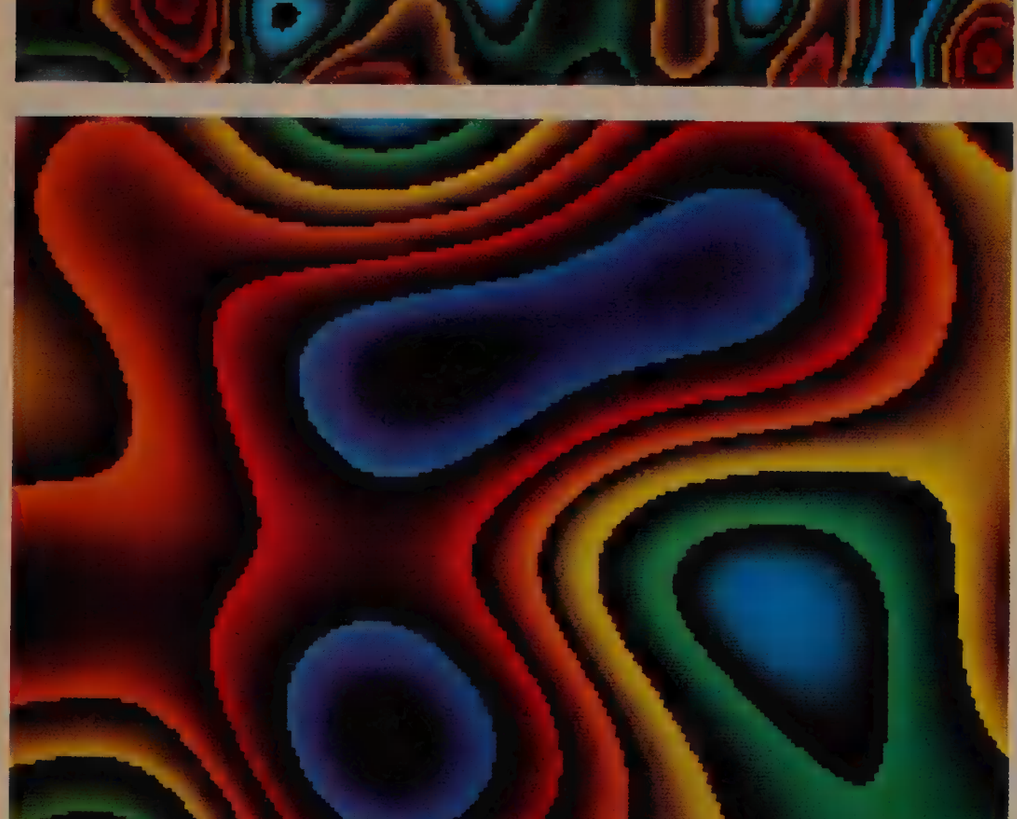
The pancake theory in its present form is a tale of two objects at the extremes of physical scale. One is an astronomical system large enough to fill 10^{23} cubic light-years of space; the other is the neutrino, a weakly interacting elementary particle that is almost vanishingly small. If the pancake theory is to be confirmed, both objects must be observed and a nonzero mass must be assigned to the neutrino. Since the two masses, if they both exist, span 80 orders of magnitude, extraordinary procedures are needed to measure them from our own vantage of middle dimensions.

Remarkably, the existence of the required scale of the astronomical system has recently been verified, and its mass offers tantalizing evidence that the pancake theory is on the right track. Systematic measurements of distance for several thousand galaxies have been carried out by determining the red shift in their spectra: the displacement of spectral lines toward the long-wavelength end of the electromagnetic spectrum. The red



DISTRIBUTION OF 400,000 GALAXIES across 100 degrees of the sky is mapped by a computer program designed to enhance the filamentary structure that is perceptible in other maps. The number of galaxies within each pixel, or small square area of the map, is indicated by a color code: black pixels represent areas having the least number of galaxies, and an increasing density of galaxies is indicated by increasingly lighter shades of brown through white. Green and red pixels correspond to local ridges and peaks in the distribution. The green pixels designate regions where the galaxy count is greater than it is in all the adjacent pixels in two or three directions, either horizontal, vertical or diagonal. Red pixels designate regions where the count reaches a local maximum in all four directions. The green filamentary ridges may correspond to superclusters of galaxies in space; if they do, they extend across 100 million light-years. The map is based on a survey made by C. Donald Shane and Carl A. Wirtanen of the Lick Observatory. Its center is the north pole of the galaxy, in the constellation Coma Berenices; its edge is the galactic parallel at 40 degrees north, and the map is oriented so that the center of our galaxy is toward the top. The map was constructed at Princeton University by John E. Moody, Edwin L. Turner and J. Richard Gott III.





SMALL PERTURBATIONS IN THE DENSITY of matter and energy were present throughout the universe shortly after the start of the big bang. The perturbations can be understood as wavelike fluctuations of the density around an average value, randomly distributed over all wavelengths. A snapshot of such a random fluctuation is shown in cross section in the upper half of this computer-generated image. Yellow, green and blue mark relatively compressed regions; orange, red and purple mark regions that are increasingly rarefied. In the lower half of the image all fluctuations encompassing a mass of less than 10^{16} times that of our sun have been filtered out by the interaction of matter and radiation in the early stages of the universe. According to the "pancake" theory proposed by the authors, the remaining fluctuations gave rise to galaxy superclusters and intervening voids that are now observed. The computer simulation was carried out by S. Djorgovski of the University of California at Berkeley.

shift is a Doppler effect, caused by the recession of a distant galaxy from our galaxy. The velocity of the recession can be calculated from the red shift according to a simple mathematical formula, and the distance of a galaxy varies directly with its recessional velocity because the universe is expanding. A redshift measurement combined with the coordinates of a galaxy on the surface of the sky enables the astronomer to fix the galaxy in space. Three-dimensional maps of the distribution of galaxies have thereby been worked out.

The maps show features quite unlike those of most other astronomical objects: the galaxies are concentrated in enormous sheets and filamentary structures whose greatest dimension, roughly 100 million light-years, is an order of magnitude larger than its lesser dimensions. Such a structure can include as many as a million galaxies; its mass is on the order of 10^{16} suns. Moreover, within each structure the galaxies are not evenly distributed: one can distinguish more densely populated clumps and strings, many of them at the intersection of two sheets. Finally, interspersed among the largest structures are huge voids, virtually free of galaxies, that are between 100 and 400 million light-years across [see "Superclusters and Voids in the Distribution of Galaxies," by Stephen A. Gregory and Laird A. Thompson; SCIENTIFIC AMERICAN, March, 1982]. Much of this picture is based on the work of Jaan Einasto of the Tartu Observatory in the Estonian S.S.R.

The detection of a massive neutrino is much more problematic. Several years ago theoretical physicists assigned to the neutrino a rest mass of zero, but some more recent theories of elementary particles suggest the neutrino does have a small mass. Several kinds of experiment are under way seeking to detect it. In the most direct method the mass can be inferred if certain variations are found in the decay rate of radioactive isotopes. In 1980 Valentin Lubimov, Evgeny Tretyakov and their colleagues at the Institute of Experimental and Theoretical Physics in Moscow measured the decay rate of tritium, the radioactive isotope of hydrogen. At that time they reported results consistent with a small but nonzero neutrino mass, between 14 and 46 electron volts, which is less than a tenthousandth the mass of the electron. Recently the same investigators have confirmed their findings and narrowed the limits of error: they now report a neutrino mass of from 20 to 40 electron volts.

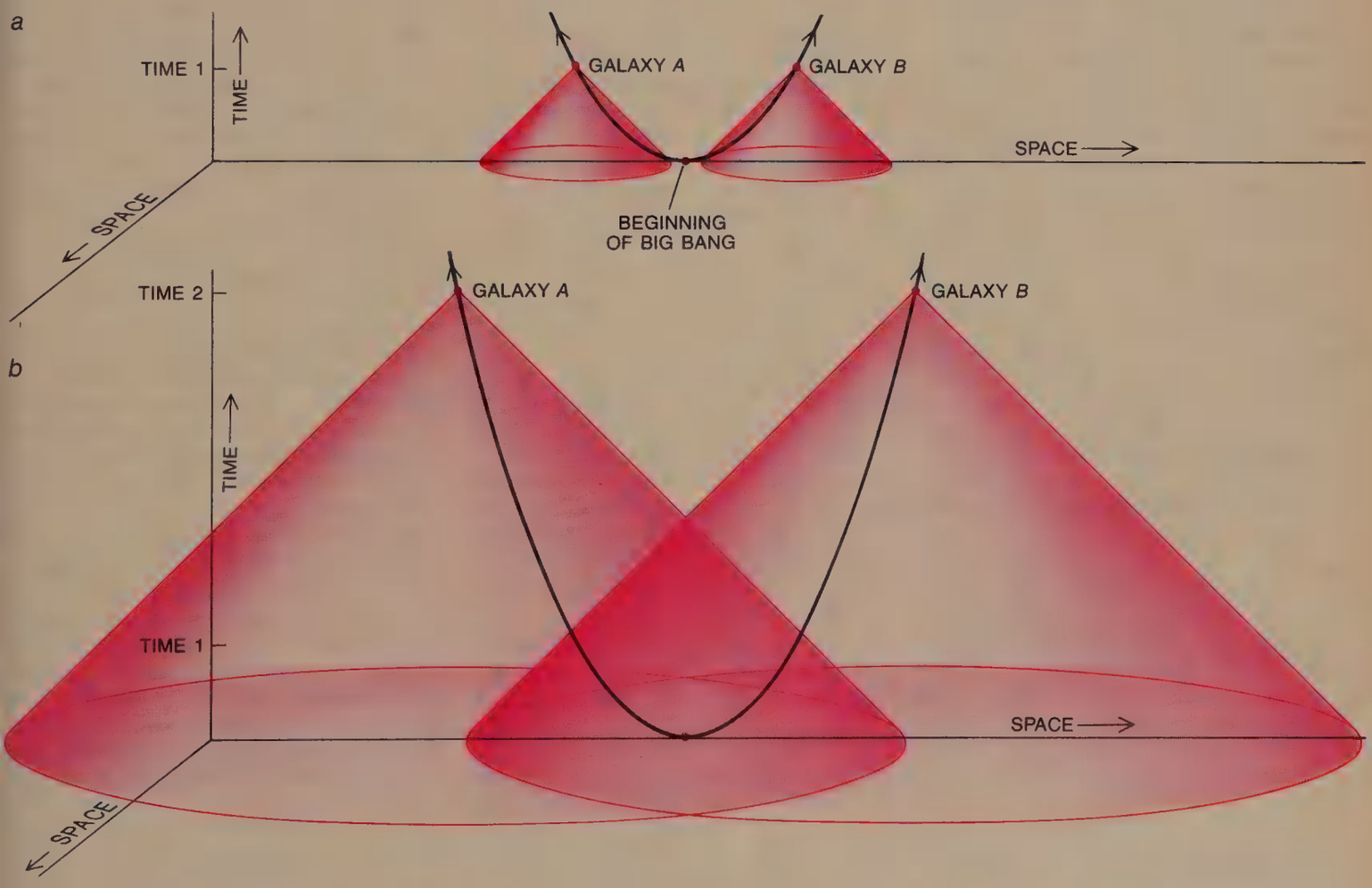
Unfortunately there has been no independent verification of these results, and so there remains no general consensus on the question of neutrino mass. A second kind of experiment, pioneered by Ettore Fiorini of the University of Milan, is based on the rate of a

mode of radioactive decay called double-beta decay that is observed in certain isotopes. Fiorini has reported the neutrino mass can be no greater than 10 to 20 electron volts, based on the decay rate of the isotope germanium 76. The method is less direct than the measurement of tritium decay; the results of Fiorini's experiment can be interpreted as a measure of neutrino mass only if it is assumed that the neutrino is its own antiparticle. On the other hand, if the neutrino and the antineutrino are distinct, the double-beta decay of germanium 76 is modified and a value for the neutrino mass cannot be inferred.

A third method of detecting neutrino mass was first proposed by Bruno M. Pontecorvo of the Joint Institute of Nuclear Research at Dubna in the U.S.S.R. The method exploits the fact that there are three kinds of neutrino: the electron neutrino, the muon neutrino and the tau neutrino. If the three kinds of neutrino have mass, if the three kinds can appear with varying probability and if the difference between the squares of the masses of any two kinds of neutrino is not equal to zero, quantum mechanics implies that the three kinds of neutrino could oscillate, or freely exchange their identities. Since the oscillations would cause the population of one kind of neutrino to vary with time, the oscillations should be detectable as a change in the population of, say, electron neutrinos along the path of a neutrino beam. Several such experiments have been done in the past few years, first in 1980 by Frederick Reines and his colleagues at the University of California at Irvine and later by Felix H. Boehm and his colleagues at the California Institute of Technology and by other workers. At this writing no experimental group has reported unambiguous evidence for neutrino oscillations. Unfortunately the absence of oscillations could merely indicate that the difference between the squares of the masses of two kinds of neutrino is zero; a failure to detect oscillations is therefore consistent with a finite, or nonzero, neutrino mass.

The prevailing attitude among physicists is that the experimental results do not yet warrant a firm conclusion about the mass of the neutrino. Nevertheless, if the evidence for mass is accepted, the cosmological consequences are far-reaching. More than a decade ago, following an early suggestion by Semyon Gershtein of the Serpukhov Institute of Physics in the U.S.S.R. and one of us (Zel'dovich), György Marx and one of us (Szalay) at Eötvös University suggested that massive neutrinos could make a dominant contribution to the mass and evolution of the universe as a whole. This suggestion was made concurrently by Ramanath Cowsik and John McClelland of the University of California at Berkeley. A massive neutrino would also lead inevitably to pancakelike structures on large scales. Before we discuss this effect, however, it will be useful to describe an earlier version of the pancake theory, a theory that ultimately failed certain critical observational tests but that has given rise to the current more successful theory.

Astrophysicists believe they have a fairly sound understanding of the physical processes that must have taken place after the first few milliseconds of the big bang. The energies of particles colliding with one another at that time



HORIZON OF AN OBSERVER grows with time and encompasses an increasingly great fraction of the universe. The horizon is a sphere, centered on the observer, whose radius is equal to the distance light can travel since the beginning of the big bang. Here the horizon is a circle at the base of a cone; the expansion of the universe since the beginning of the big bang is indicated by the divergence of two galaxies with time. In the early stages of expansion the galaxies recede from each other at an apparent velocity that exceeds the speed of light, and so at this epoch there has not been enough time for an observer in one of them to see the other galaxy at any previous stage (*a*). Because the expansion is slowing, however, light from an early stage of one of the galaxies will eventually reach an observer in the other galaxy (*b*). The edges of the cones are the paths of light signals in space; since no signal can propagate faster than light, they represent a spatial limit to causal interaction at any given time as well as a limit to observation. Hence any fluctuation in the density of the universe on a scale larger than that of an observer's horizon will have no effect on the observer or on the distribution of matter and energy within the horizon.

were no greater than the energies typically achieved in small particle accelerators, and so a picture of the early universe emerges when one considers a dense fluid of particles whose individual properties are well known.

By far the most abundant particles in the fluid were the photon, the electron and the three species of neutrino; only relatively small numbers of protons and neutrons were left over from annihilations that took place in an earlier epoch. The electrons and the neutrinos remained in close contact throughout most of the first second and were continuously created and annihilated. Frequent collisions among them guaranteed that energy was distributed randomly throughout the fluid; in other words, the particles were maintained in thermal equilibrium. As the universe expanded, the density of the particles decreased and collisions became less frequent. Because the energy of a photon varies inversely with its wavelength, the average energy of the photons decreased as their wavelengths expanded with the rest of the universe, and so the universe began to cool.

Recent theoretical investigations that have sought to unify the fundamental forces of nature can now peer even farther back than the first millisecond into the history of the universe. The theories are called grand unified theories because they attempt to understand the electromagnetic force, the weak nuclear force and the strong nuclear force as distinct low-energy manifestations of a single, underlying phenomenon. (Gravity, the fourth fundamental force, has not yet been incorporated.) The energy density at which the three forces become indistinct corresponds to the energy density of the universe only 10^{-35} second after the start of the big bang. The early universe has therefore come to be regarded as a laboratory for testing the predictions of grand unified theories.

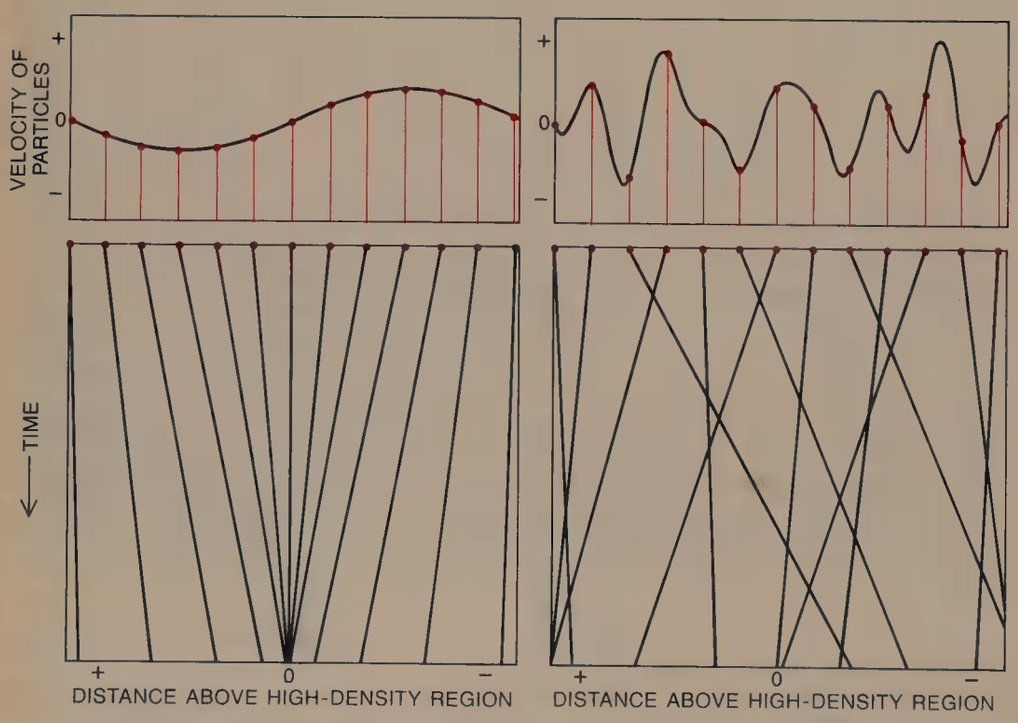
One prediction of the theories is that if the density of matter must fluctuate in the early stages of expansion, the density of the photons, or radiation, must fluctuate also. Nevertheless, the ratio of the density of matter to the density of radiation must always remain the same. According to the general theory of relativity, energy and matter are equivalent as a source of gravity and determine the geometry of space-time. A fluctuation in the density of mass and energy therefore causes a fluctuation in the gravitational field, which is equivalent to a fluctuation in the curvature of space-time. The comprehensive theory of such fluctuations in the expanding universe, treated within the framework of the general theory of relativity, was developed in 1946 by Eugene M. Lifshitz of the Institute for Physical Problems in Moscow.

It seems reasonable to assume that fluctuations must have existed in the early universe over a wide range of possible scales. We make the assumption primarily for reasons of parsimony: it would seem arbitrary and entirely fortuitous if the initial fluctuations were to single out regions only of, say, galactic scale. There is an upper limit, however, to the size of fluctuation that can be perceived by any observer at a given time. That limit is the spatial horizon of the observer, which is a sphere, centered on the observer, whose radius is equal to the distance light can travel in the time since the start of the big bang. In the standard model of the big bang, however, the initial expansion of space and time from the singular point creates a universe far greater than the spatial horizon of any single observer. On the other hand, because the expansion of the universe is thought to be slowing down, increasing amounts of mass come within any observer's horizon. Fluctuations undetectable in the early universe become detectable later on, because they begin to be encompassed by the observer's ever widening horizon [*see illustration on preceding page*].

Once a fluctuation is within an observer's horizon it is adequately described by classical, or nonrelativistic, gravitational theory. There it takes the form of an observable perturbation in the density of the fluid. There are two competing effects on any parcel of matter and radiation moving together: gravity tends to collapse the parcel, and pressure caused by the chaotic motion of the particles and radiation tends to disperse the parcel in space. On large scales gravity always wins. Pressure cannot resist the collapse, so that particles are attracted to the regions of highest density. Moreover, once gravitational collapse begins, the accreted mass can attract more distant mass and radiation, and so any initial instability is amplified. Matter accumulates in some regions and becomes rarefied in others.

If the soup of particles and radiation that make up the early universe is considered to be an ideal gas, the effects of a density fluctuation superposed on the gas are straightforward. Any local compression in the density over a sufficiently large mass will trigger gravitational instability and lead to incipient collapse. On smaller scales, however, gravity is not strong enough to overcome the increase in the pressure of the gas caused by the increase in density. The compressed parcel of gas will therefore rebound and become rarefied, and the fluctuation will propagate exactly like a sound wave, that is, by the periodic compression and rarefaction of the medium through which it travels.

Most sound waves in the air die out in a few tens of meters because the particles that make up the pressure waves are scattered and their coherent motion is dissipated as heat. In a similar way sound waves in the cosmic medium set up by the fluctuations lose their energy and die away at all but the longest wavelengths. Furthermore, the particles and photons in the early universe are much



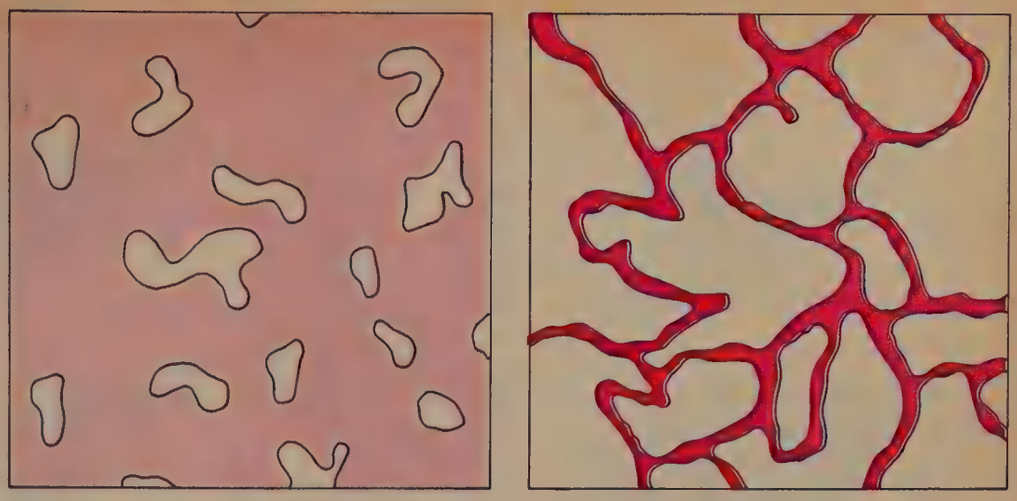
SMOOTH DISTRIBUTION OF PARTICLE VELOCITIES arises from a fluctuation with a single wavelength; particles are gravitationally attracted to the regions of highest density, and so the velocity of a particle depends on its distance from such a region, as is shown in the graph at the upper left. In the graph at the lower left the velocity of each particle is given by the slope of the line that represents its trajectory. The trajectories tend to converge and form regions of enhanced density. Thus the smooth fluctuations are amplified. If the fluctuations are random over all the wavelengths (*upper right*), the trajectories do not converge (*lower right*).

too densely packed to be treated as an ideal gas. In the first 300,000 years of the big bang the photon radiation was energetic enough to keep all matter ionized. Photons outnumbered electrons by a factor of about 100 million; the free electrons, which would later be bound to atomic nuclei, were therefore under constant bombardment by the photons, freely scattering the photons and being scattered by them. The result was a thick, viscous fluid of electrons and photons in which no particle could travel very far without being scattered.

The scattering of the free electrons by the radiation makes any displacement of the electrons through the radiation much like moving through viscous, cold molasses. The viscosity of the fluid thereby inhibits the growth of gravitational instabilities that might be caused by the accretion of matter alone. Moreover, as in the ideal gas, the large pressure of the radiation keeps matter and radiation from collapsing together under the pull of gravity on all but sufficiently large scales. The remaining fluctuations within the viscous fluid, that is, the ones that survive gravitational collapse, can be regarded as sound waves.

As we have mentioned, grand unified theories require that the ratio of the density of photons to the density of matter remain always the same: in the compression phase of a fluctuation the compression of the photons must therefore match the compression of the particles with mass. If the distance covered by a photon in the time since the start of the big bang is greater than the distance across a compressed region of a fluctuation, however, the photon will in effect not take part in the compression but will instead dissipate its share of the energy of the fluctuation. Since the photons greatly outnumber the particles with mass, they carry almost all the energy of the fluctuations, and so fluctuations on a scale smaller than the average radial displacement of a photon in the time since the start of the big bang are damped out.

The path of a photon can be compared to the path of a drunkard staggering away from a lamppost along any direction with equal probability. In order to wander a distance away from the lamppost that corresponds to N steps when sober, the drunkard must take N^2 steps. Similarly, the photon must be scattered N^2 times in order to travel a radial distance equal to the distance it would travel if it were streaming freely. In spite of their being scattered by the electrons, the photons diffuse radially outward through the medium so fast that they dissipate the energy of all but the largest fluctuations. By the time the universe cools enough for atomic nuclei to capture the free electrons, the photons have diffused through a region of the universe whose mass is about 10^{14} times that of our sun. All initial fluctuations that encompass a mass smaller than this amount are therefore erased.



TOPOLOGICAL STRUCTURE OF THE FLUCTUATIONS that survive filtering in the early stages of the universe is preserved in the structure of galaxy superclusters and voids. At the left the density of the universe is nearly uniform. If the probability of rarefaction caused by a slight fluctuation in the density along any one axis is one-half, the probability that any region is rarefied along all three spatial axes is one-eighth. Such regions are shown in white; they initially occupy only one-eighth of the volume of the universe and include about one-eighth of its mass (*left*). Gravity compresses the colored regions along one or more axes. These regions, which include about seven-eighths of the mass of the universe, collapse to form a cobweb of filaments that eventually fill only one-eighth of the volume of space (*right*). The regions of initial rarefaction expand to make voids that fill the remaining seven-eighths of the universe.

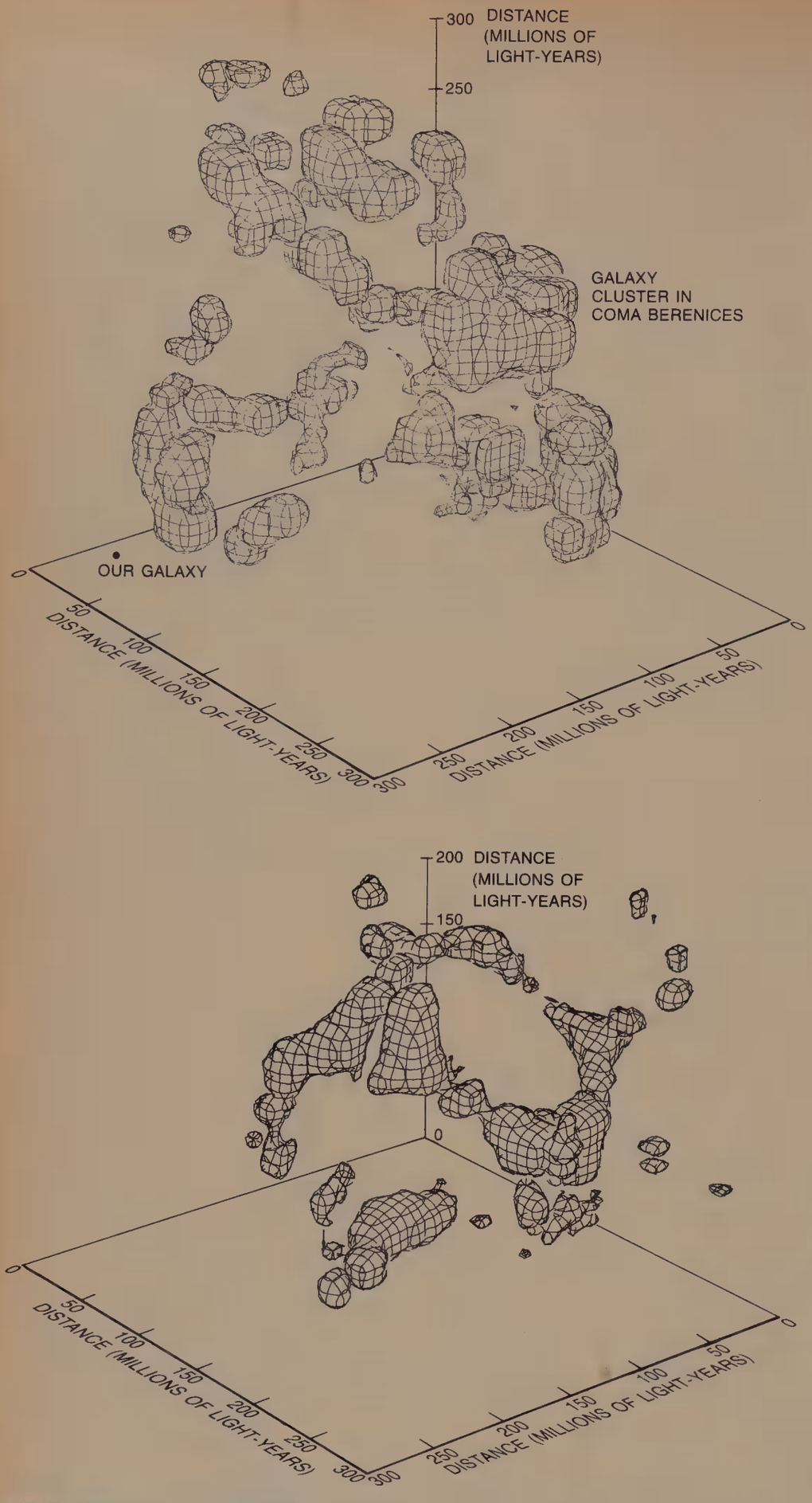
When the electrons finally combine to form atoms, the matter and the radiation cease to interact and the radiation streams off independently of the atoms. The viscosity of the fluid drops abruptly, and the fluctuations that have survived the previous era of radiation-dominated interactions are no longer inhibited from being amplified. Thereafter gravitational instability proceeds with full vigor.

The sudden absence of radiation pressure has a dominant effect in determining the shape and structure of the first objects to form. Thermal pressure always acts isotropically, or equally in all directions, and so if the radiation pressure had remained comparable in strength to gravity, all collapsed objects would have had nearly perfect spherical symmetry. Anisotropies develop because the pressure is completely negligible up to the last moments of collapse.

Because of the lack of pressure to counteract the infall, gravitational instability is highly efficient at sweeping almost all the matter into compressed, high-density regions of space. Consider the following argument. Along any one of the three spatial axes matter can be either compressed or rarefied; assume, for the sake of simplicity, that the probability the matter is compressed along any one axis is one-half. The fraction of the gas that will not be compressed along any axis is the cube of one-half, or one-eighth. This result has immediate implications for the predicted spatial structure following collapse. At an early stage, when the density is still nearly uniform, the regions destined to be compressed include about seven-eighths of all matter. These regions surround smaller bubbles of matter that never collapse; the bubbles are destined to become voids. After the collapse the compressed regions occupy only one-eighth of the volume of space; the small bubbles, which carry one-eighth of the matter, expand to fill the remaining seven-eighths of the volume. The topology of the initial state is preserved. The final outcome is a cellular structure formed by thin walls and filaments of compressed matter that enclose huge voids.

The shape of the compressed regions can be predicted from similar considerations. It is most unlikely that any cubical volume of matter destined for collapse will form a sphere. Such a collapse would require a match of both direction and magnitude of the fluctuations along all three components into which any arbitrary collapse can be resolved. It is much likelier that the cube would collapse first along one randomly selected axis, and it would collapse or expand more slowly along the other two axes. The ensuing distribution of matter is highly anisotropic. Since the mass inside the initial cubical volume does not change as both the thickness and the volume of the cube decrease, the density becomes extremely high and a flat pancake is formed.

At first the pancakes develop in isolated regions, but they soon grow into thin sheets that intersect and form the cellular structure. Numerical simulations of the collapse done with the aid of large computers suggest the universe has only recently acquired a cellular structure. In the future, as larger clumps of matter form, the cellular structure is expected to disappear. Hence it is only during an



CONTOURS OF HIGH DENSITY are plotted in three dimensions for all galaxies brighter than magnitude 14.5 in the northern sky that are within about 250 million light-years of our galaxy (*upper illustration*). The distribution is similar to the one shown in the lower illustration, which is derived by simulating the predictions of the pancake model on a computer. In the model it is assumed the clustering of matter arises from density fluctuations no smaller than the smallest fluctuation to survive the smoothing effects of hypothetical massive neutrinos. The real distribution was constructed by Carlos Frenk and Simon White of the University of California at Berkeley from a map prepared by Marc Davis and his colleagues at Center for Astrophysics of the Harvard College Observatory and the Smithsonian Astrophysical Observatory. Map was based on a survey of 1,801 galaxies. Simulation was done by Frenk, White and Davis.

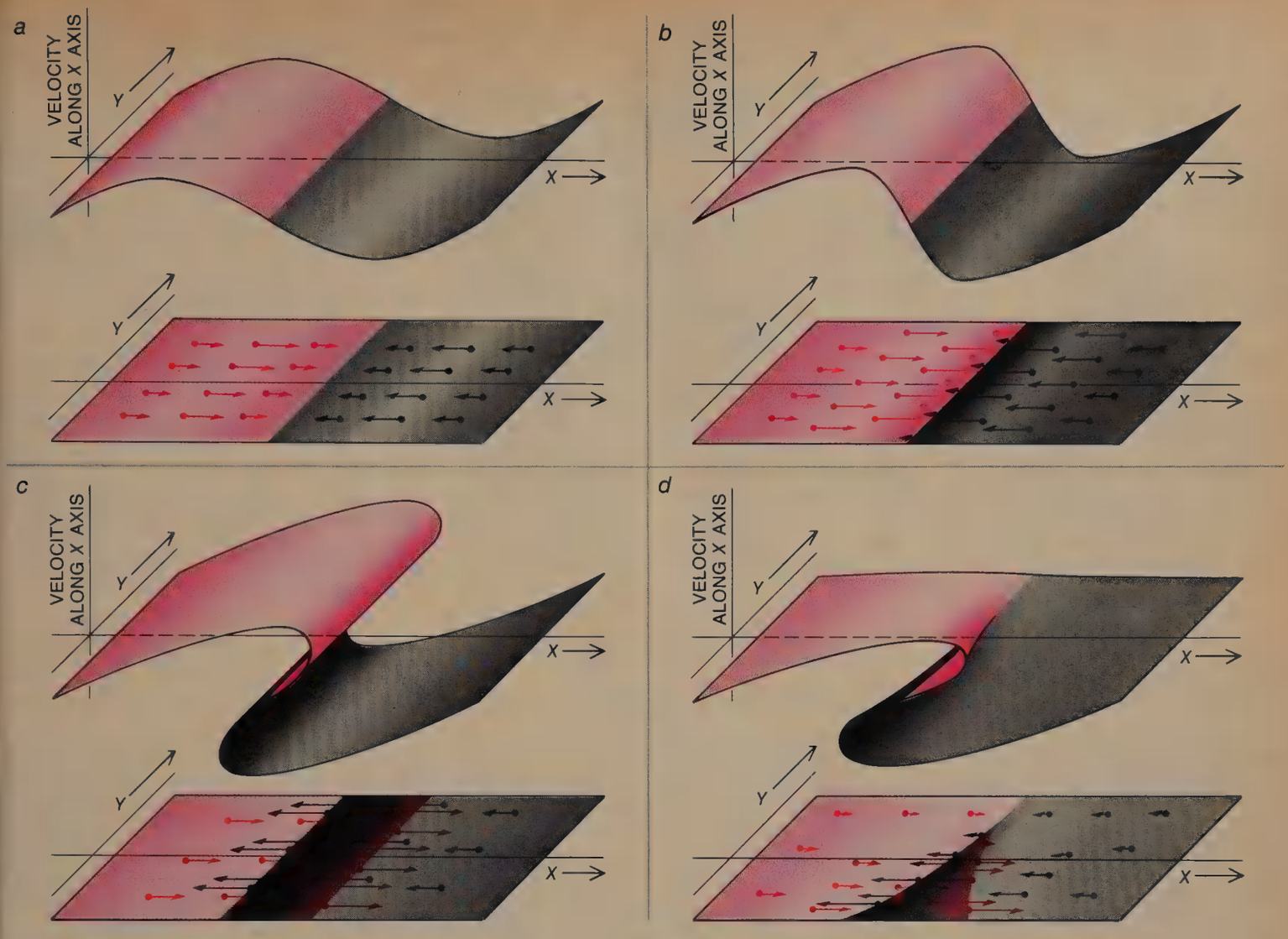
intermediate stage of cosmic evolution that the initial curvature fluctuations are reflected by the structure of matter. The observational evidence shows that from the perspective of the evolution of large-scale structure the universe is neither very young nor very old.

There are two major difficulties with the pancake theory as we have described it so far. First, remember that the smallest fluctuations to survive the radiation era encompass a mass of 10^{14} suns. Structure in the distribution of galaxies, however, exists at much larger scales. Numerical simulations favor a theory in which the smallest fluctuations emerging from the radiation era are on a scale of 10^{15} or 10^{16} suns.

The second difficulty is more serious. Because the microwave background radiation has propagated freely ever since the photons and the electrons ceased to interact, the variation in the temperature of the radiation across the sky reflects primordial inhomogeneities in the distribution of matter. At the time the original pancake theory was formulated the upper bound for the temperature variation over the entire sky was about one part in 1,000. Accordingly it was thought that matter inhomogeneities in the early universe could have been as great as a third of the temperature variation, or one part in 3,000. Recently more stringent limits on the variation of the radiation temperature have been set by Francesco Melchiorri and his co-workers at the University of Florence and the University of Rome, and by Yuri N. Parijskijin of the Pulkovo Observatory in Leningrad. The new upper bound is a variation of less than one part in 30,000 over an angle of six degrees.

The fluctuations required by the original version of the pancake theory were consistent with the earlier estimate of temperature variations, but the agreement with the new estimate is only marginal. Moreover, if the overall density of matter and energy in the universe is so small that the present expansion will continue forever, the agreement is lost. On a cosmic scale the force of gravity would have been so weak in recent epochs that fluctuations must have completed their growth and collapsed at a much earlier time, when the density of matter was much greater than it is today. The amplitude of such fluctuations, however, would have been much too large to be compatible with the uniformity in the microwave background. On the other hand, if the initial fluctuations had been small enough to be compatible with the radiation background, the birth of galaxies would have become practically impossible.

If the universe is dense enough for the amplitude of the fluctuations to be marginally reconciled with the uniformity of the background, another problem



DISCONTINUITIES IN THE DENSITY OF MATTER can result from continuous deformations of the medium during gravitational collapse. In each frame of this schematic sequence the lower plane represents two directions in space; for simplicity only the motions of particles in the plane and parallel to the *x* axis are considered. The particles move toward a central axis of the plane; the length of each arrow indicates the velocity of a particle at the position corresponding to the tail of the arrow (*a*, *b*). A sheet of particles moving to the right (*colored arrows*) crosses the central axis without colliding with the particles moving to the left (*black arrows*). The motions set up two discontinuities in the density of the medium, one on each side of the central axis (*c*). In almost any real collapse the motions of the two sheets of particles are not exactly symmetrical and the two discontinuities end in a cusp (*d*). The same density distributions can be understood as a special case of a more general phenomenon described by the branch of mathematics called catastrophe theory. If the motions of the particles are plotted in a phase space, that is, in a three-dimensional space where the vertical axis represents the velocity in a direction parallel to the *x* axis, the interaction of the two sheets of particles is represented by a twisting or undulating surface (*upper plane in each frame*). The density of the particles at any point is then given by the "shadow" cast by the deformed surface on the original *x*-*y* plane.

arises. The density cannot then be accounted for solely by the total mass of bright matter, visible as stars, nebulas, galaxies and the like. Instead the universe must be made up predominantly of dark matter. This inference is not a new one. Studies of the rotation of our galaxy and that of other spiral galaxies have shown that the rotational velocities of stars on the periphery of a galaxy are not consistent with Kepler's laws. These laws state that the rotational velocity should decrease with increasing distance from the center of a galaxy, just as the orbital velocity of a planet decreases with its distance from the sun. Peripheral stars, however, do not slow down; their rotational velocities are roughly constant and independent of their distance from the galactic center. P. James E. Peebles and Jeremiah P. Ostriker of Princeton University and Einasto simultaneously suggested the dilemma would be resolved if halos of invisible matter make up the bulk of the mass of spiral galaxies. An indirect argument suggests dark material may be present in even greater quantities within groups and clusters of galaxies. Such systems would fly apart in an unaccountably short time if it were not for the gravitational attraction of dark matter. It is estimated that dark matter may comprise 90 percent of the mass of the universe.

A new component of the universe was badly needed to salvage the pancake theory, and a source of dark matter was needed to account for the motions of galaxies. A natural candidate for both purposes was the neutrino, although certain other exotic but still undetected particles, such as a massive photino or a massive gravitino, might serve the same cosmological function. Theories of elementary particles predict that in the first millisecond of the big bang a wide variety of weakly interacting particles were maintained in thermal equilibrium. Many such particles could still survive, and provided they are stable they could have far-reaching implications for cosmology. Since the neutrino mass can be measured experimentally, in the remainder of this discussion we shall refer to the neutrino. Nevertheless, even if the neutrino turns out to have no mass, the pancake theory is not disproved.

Remember that in the first second of the big bang the primordial soup included an abundance of neutrinos. Even today the ratio of photons to all three varieties of neutrino is only 11 to 9. Neutrinos, unlike protons, electrons and even photons, interact so weakly with other particles that they begin to stream freely through the fluid long before the photons do. Hence the neutrinos, which initially move at the speed of light, can



COMPUTER SIMULATION shows the evolution of large-scale fluctuations in the density of matter and energy with time, assuming the existence of neutrinos that have a nonzero mass. Pancakes and filaments develop as a result of gravitational collapse, and the matter in the remaining regions of space becomes increasingly rarefied. The resulting structure resembles the distribution of superclusters and voids that is currently observed. In several regions there are cusps and other kinds of discontinuity in the density, which can be recognized and classified by means of catastrophe theory. The simulation was done by George Efstathiou of the University of Cambridge.

travel farther than the photons in the early stages of the universe. By the end of the radiation era the neutrinos have dissipated fluctuations on a larger scale than photons alone could have done.

A massive neutrino cannot continue indefinitely moving at the speed of light. When the energy density of the photons falls below the energy that corresponds approximately to the rest mass of the neutrino, the neutrino begins to slow down and move at a speed appropriate to its energy. If the mass of the neutrino is 30 electron volts, the slowdown will begin well before the capture of the free electrons by atomic nuclei. The capture must wait until the background energy is reduced to .1 electron volt, the energy at which hydrogen is ionized by the dense fluid of photons. Although the neutrinos continue to erase the fluctuations as they slow down, they become increasingly susceptible to being trapped by large fluctuations that have not yet been smoothed out. Richard Bond of Stanford University and one of us (Szalay) have estimated the maximum scale over which the neutrinos can freely stream before they are trapped, and consequently the minimum scale over which the fluctuations are not erased. The scale corresponds to a present distance of 100 million light-years and a mass of 10^{15} to 10^{16} suns. The agreement with the size and mass of the galaxy superclusters that are now observed is striking.

How can such fluctuations be compatible with the observed uniformity of the background radiation? The neutrinos cease their erasure of the curvature fluctuations before the end of the radiation era, but unlike the electrons their motions are not inhibited by the viscosity of the fluid. Neutrinos collide so rarely with photons or electrons that they are not subjected to viscous drag. Gravitational instabilities among the neutrinos can accordingly begin to develop before the end of the radiation era, and so they can grow over a much longer time than the fluctuations of ordinary matter can. The initial amplitude of the neutrino fluctuations needed to account for the present inhomogeneities of matter could therefore have been much smaller than the amplitude of the fluctuations needed in a mixture of radiation and ordinary matter. With massive neutrinos the variation in the temperature of the background radiation required to generate the observed aggregations of matter is reduced by an order of magnitude or more. Thus theory and observations can be reconciled.

The new version of the pancake theory leads to a natural explanation for the origin of the dark matter in the universe. The initial collapse of a pancake distributes most of the neutrinos widely because most of them are accelerated by the collapse to large velocities, on the order of 1,000 kilometers per second. Such neutrinos are destined to fill the dark regions of intergalactic space. Other neutrinos, however, move more slowly because they are initially closer to the central plane of the pancake and do not undergo large accelerations. The thin layer of gas in the vicinity of the central plane condenses and breaks up to form protogalaxies. The slow-moving neutrinos are gathered together by aggregates of ordinary matter, and the matter near the center of the protogalaxy continues to condense and form stars. Neutrinos at the periphery of the protogalaxy, however, are gravitationally shared and never condense; they become the dark matter in the galactic halo.

A more detailed theory of galaxy formation within the context of the new pancake theory is now under development. As a pancake collapses, the neutrino component of the collapsing gas passes through the central plane of the pancake without interaction. The density distribution of the neutrinos acquires sharp discontinuities, some of which can be identified with rich clusters of galaxies. Vladimir I. Arnold, a mathematician at Moscow State University, has recently collaborated with astrophysicists on the problem, and he has identified such discontinuities in the overall density distribution with certain elementary structures in the branch of mathematics called catastrophe theory.

The pancake theory, as it has now been modified, offers deep insight into the character and origin of the contemporary structure of the universe. It is based on well-known physical principles and on plausible assumptions about the conditions in the very early universe. As a theory of the origin of large-scale structure it is by no means unique, although it now appears that both theory and observation point in the general direction we have outlined. Nevertheless, there are many important questions that must be resolved before the theory can be considered firmly established.

Given the confirmation of the theory, there are two main lines along which it must be developed. First, one must address the finer structure of the universe, the formation of the first generation of stars from a primordial gas that was totally devoid of the heavy elements. Second, one must ask how the initial conditions assumed by the pancake theory arose from even earlier epochs in the history of the universe. Attempts are now in progress to show how the small-amplitude fluctuations required by the pancake theory could result from earlier and much less well-understood phenomena. These attempts are based on theories not yet settled, but preliminary results leave room for optimism that by the end of the century we shall possess a coherent theory of the universe.

Reshaping the future of film.

From fat to flat. The goal many of us strive for with laborious push-ups and strenuous diets, scientists have achieved in Kodak laboratories. The "bodies" they have streamlined, however, are those of silver halide grains. More than a beauty treatment, this process of reshaping enabled us to position grains for maximum light-catching ability, and expand their sensitizing characteristics. It's this advance in silver halide emulsion technology which made it possible for us to produce a true 1000-speed film with a balance of speed, grain, and sharpness previously unattainable.

The structure of new Kodacolor VR 1000 film is a modification of the inverted-layer format first used in Kodacolor 400 film. This concept places the fastest light-sensitive layers closer to the camera lens. The magenta and cyan layers are repositioned to form a fast-magenta, fast-cyan, slow-magenta, slow-cyan sequence. In the new emulsion, the yellow-filter layer is eliminated, and the protective overcoat and ultraviolet-filter layers are combined into one, without any loss in color quality.

The resultant faster film can be exposed by daylight, electronic flash, or tungsten, which provides photographers with greater picture-taking capability, particularly when recording fast-action scenes or shooting in low-light conditions.

While reshaping the light-sensitive substance of film is a significant accomplishment in itself, we view this new technology as a stepping-stone. It may well reshape the future of film and enable photography to do hundreds of different jobs better than before.

For more information on Kodak's fastest color print film ever and the new emulsion technology behind it...send your request to: Eastman Kodak Company, Department GBSA-8, 343 State St., Rochester, NY 14650.

© Eastman Kodak Company, 1983

Kodak

Kodak. Where technology anticipates need.



SCIENCE AND THE CITIZEN

End of the Chain

The worldwide recession that began in 1980 is beginning to ease after three years (making it the longest economic slump since the depression of the 1930's), but one of its legacies has been a serious setback for worldwide development. There is a chain of cause and effect, with the poorest countries at the end of the chain. Recession in the advanced industrial countries has reduced demand for third-world products. It has also increased unemployment, encouraging protectionism and further thwarting developing countries' efforts to export. The terms of trade have turned against the developing countries: inflation has raised the cost of industrial products they need to import while the prices of commodities, their primary exports, have declined. Having turned increasingly to intergovernmental and commercial borrowing as a source of the capital they need, the developing countries have been hit by rising interest rates on the money they owe, a result of the restrictive monetary policies adopted by the industrial countries in an effort to counter inflation.

The foregoing picture is sketched out by the World Bank's *World Development Report* for 1983. According to the report the recession of 1980–82 was the second in less than 10 years. Growth rates of output in the industrial economies and of world trade fell for two years after the oil-price jump of 1973 but recovered rapidly in 1976; the developing countries were not seriously affected. The recent recession has lasted longer, in part because of tightened monetary controls. With monetary restraint accompanied by large fiscal deficits, interest rates rose, reaching a peak in 1981 and prolonging the recession. Unemployment in the industrial countries, which the World Bank considered "high" at about 5 percent after the first recession, climbed to 8 percent in 1982 (and to almost 11 percent in the U.S.). The recession slowed the growth of world trade; exports actually declined in 1982.

The industrial countries influence economic growth in the developing world primarily through international trade and the flow of capital. The recession reduced both total demand and the price of typical third-world exports: foods and industrial raw materials. Adjusted for the inflated cost of manufactured goods the developing countries' import, commodity prices were lower in 1982 than at any time since World War II. Many of the poorest countries, which rely heavily on a single export commodity, were particularly hard hit. Developing countries that had managed to increase their exports of manufactured goods came up against a surge of protectionism in the form of tariffs, import quotas and other restrictive measures. Export earnings of developing countries fell in 1982 to 7 percent below the 1980 level.

The large gap between receipts from exports and the expenditure of foreign exchange for imports and debt servicing is financed in most developing countries largely by official development assistance ("foreign aid") and new borrowing. Development assistance (both loans at low interest rates and grants) comes primarily from the members of the Organization for Economic Cooperation and Development (OECD) and of the Organization of Petroleum Exporting Countries (OPEC). The former have accepted a "target" of .7 percent of their gross national product as their annual contribution to development assistance; few have reached the target. Whereas the Netherlands, Norway and Sweden devoted more than 1 percent of their GNP to such aid in 1982, the percentage was .48 for West Germany, .29 for Japan, .27 for the U.S. and .25 for Switzerland. OPEC has done better, but total development assistance has not increased since 1980.

With exports down, official assistance insufficient and foreign-exchange reserves declining, third-world officials increasingly have turned to private lenders, who are now owed more than 70 percent of the developing countries' total debt. In 1982 developing countries had to pay nearly $50 billion in interest on medium- and long-term debt, 50 percent above the 1980 level. Their total debt came to more than their total earnings from exports; their annual debt-service payments were almost 21 percent of their export earnings.

According to the World Bank's report, sustained world economic growth must begin with vigorous growth in the industrial nations and an expansion of world trade. The report emphasizes that improvement in third-world exports is necessary not only for humanitarian and political reasons but also in the self-interest of the industrial nations. Most obviously, borrowing countries can service their debt only if they earn enough foreign exchange by exporting. Moreover, protectionism "is costly to the protectionist countries themselves." It raises the cost of goods they consume, to the special detriment of their poorest consumers. It hampers needed restructuring of many industrial nations' economies by preserving inefficient industries and postponing investment in new and more appropriate ones. Moreover, by restricting the growth of developing countries protectionism restricts the expansion of what has become a major market for the industrial nations. The U.S., for example, now sells two-fifths of its exports to developing countries.

Five-Year Plan

Although the construction of nuclear power plants has slowed considerably in the U.S., the industry is continuing to expand in other countries. France, for example, gets more than 40 percent of its electric power from nuclear energy and plans to reach 85 percent by the end of the century. One country that is known to have a strong commitment to nuclear power but has supplied few details of its program is the U.S.S.R. A rare glimpse of those details has now been provided in *IAEA Bulletin,* a publication of the International Atomic Energy Agency, by B. A. Semenov, a Russian who is deputy director general of the agency and head of its Department of Nuclear Energy and Safety.

Semenov writes that by the end of last year the total installed capacity of the nuclear power plants in the U.S.S.R. exceeded 18,000 megawatts and that in 1981 they generated 86 billion kilowatt-hours of electricity, or 6.5 percent of the electric power. (The Atomic Industrial Forum recently tallied 37 operating reactors in the U.S.S.R. but was unable to ascertain what fraction of the country's power they generated. The U.S. has 72 operating commercial reactors, accounting for 13 percent of the power.)

A sense of the pace of the Russian program is given by Semenov's statement that "during the next five-year period the generation figure will increase more than threefold and will reach 220 billion kilowatt-hours in 1985." (The U.S.S.R.'s current five-year plan for economic development covers the period from the beginning of 1981 to the end of 1985.) Moreover, "since it is planned that the total electric power production in the country will be increased to about 1,500 billion kilowatt-hours, nuclear plants will account for 14 percent of the total electricity production in the entire country by 1985 and for 24 percent in the European region."

Fundamental differences between the European part of the U.S.S.R. and the Asian part constitute one of two reasons cited by Semenov for the nation's decision "to favor nuclear power as a main source of energy." About 80 percent of the country's energy resources are concentrated in the eastern region, whereas 75 percent of the population and the consumers of power are in the European region. As a result the transportation of fuel from east to west constitutes about

40 percent of the country's rail-freight operations. The second reason is that "nuclear power is less damaging to the environment than conventional power."

Although the U.S.S.R. opened its first nuclear power plant in 1954, the pace of expansion has been most rapid in the past few years. Of the 18,000 megawatts in installed nuclear capacity some 12.5 million kilowatts is supplied by plants put into operation in the period from 1976 to 1982. The rate at which nuclear power was introduced nearly tripled in the last five-year period compared with the preceding one. The rate at which nuclear power plants are being introduced is about 2.5 times higher than the rate of introduction of power plants burning fossil fuels. During the current five-year period the construction of new fossil-fueled plants in the European part of the country "will practically cease," according to Semenov, "and by 1985 almost all the increase of installed capacity will be from nuclear power plants."

The U.S.S.R. also plans to give a good deal of attention to breeder reactors, which are designed to make more nuclear fuel than they consume. (The present program rests on two types of reactor: the WWER pressurized, light-water-moderated-and-cooled reactor and the RBMK light-water-cooled, graphite-moderated channel type.) Breeder reactors have "a special priority," Semenov says, because the U.S.S.R. "clearly recognizes that the solution to long-term nuclear-fuel problems for large-scale nuclear-power programs requires the wide use of breeder reactors."

The Mitochondrion's Message

The concept that aberrant nucleotide sequences in DNA can be passed from parent to child and cause inborn disorders of metabolism is now a familiar one. The nucleotide sequences carry the instructions for making enzymes and other proteins that are essential to the functions of the living cell. Therefore a disruption of the correct nucleotide sequences can lead cells to make abnormal proteins or to make none of a particular protein. In all the familiar instances of inherited human metabolic disorders, such as sickle-cell anemia and beta thalassemia, the incorrect genetic information is in the DNA of the chromosomes in the nucleus. Not all the genetic material of the cell, however, is in the chromosomes. For example, DNA is also found in the mitochondria: the numerous energy-producing organelles in the cytoplasm surrounding the nucleus. In a recent issue of *The New England Journal of Medicine* Joseph Egger and John Wilson of the Hospital for Sick Children in London report the first instance where a metabolic disorder has been shown to be transmitted from parent to child by the mitochondrial DNA.

Each mitochondrion consists of an outer membrane wrapped around an inner one to form a complex, irregular oblong. In the spaces defined by the two membranes are many copies of a large set of enzymes. The enzymes mediate the oxidation of pyruvate, fatty acids and amino acids to carbon dioxide and water. The oxidation is coupled to the synthesis of adenosine triphosphate (ATP), which serves as the fuel for many cellular reactions.

In order to carry out its function of making ATP each mitochondrion needs a full complement of enzymes. The genetic information required to make the proteins is divided between the DNA in the nucleus and the DNA in the mitochondrion. The code for most of the enzyme components is carried by the chromosomes. The code for the rest of the enzyme components is carried by the genetic material of the mitochondrion: a double-strand DNA circle that in human cells is a few thousand nucleotides long. The chromosomally coded proteins are made on ribosomes in the cytoplasm and transported into the mitochondrion. The mitochondrially coded proteins are made on ribosomes within the mitochondrion.

Since some enzymes are coded in the nucleus whereas others are coded in the mitochondrion, a genetic disorder of the organelle could have its origin in the nuclear DNA or in the mitochondrial DNA. The nuclear and mitochondrial forms of inheritance can give rise to quite different patterns of transmission.

The nuclear pattern is Mendelian inheritance. Each trait that shows such inheritance is determined by at least two genes, one on each member of a pair of chromosomes. Each gamete (sperm or egg) includes only one of the pair of chromosomes. As a result the child does not necessarily inherit the trait: it appears in a complex pattern in the next generation, with some children inheriting it and others not inheriting it. The pattern is called segregation, and if it is known whether the trait is dominant or recessive, the probability of inheritance can be calculated with considerable precision. It can readily be seen that a Mendelian trait can be inherited from either the mother or the father, because the zygote (fertilized egg) contains one set of chromosomes from each parent.

The traits coded by mitochondrial genes, however, are transmitted by a quite different mechanism. Almost all the mitochondria in the zygote come from the mother. The sperm contributes very little cytoplasm to the zygote and therefore contributes few cytoplasmic organelles or none. Hence the mitochondrial DNA comes almost exclusively from the mother.

Egger and Wilson noted the appearance of the syndrome called mitochondrial cytopathy in successive generations of 30 families; in each family the disorder was known to be present in at least two generations. Mitochondrial cytopathy entails defects in the shape of the mitochondria and in the structure of several mitochondrial enzymes. The symptoms of the disease include reduced height, muscular weakness and many malfunctions of specific organs.

In 27 of the 30 families the disease was exclusively transmitted through the mother. Fifty-one mothers and only three fathers had passed the disease to their children. Of the 130 siblings of definitely diagnosed victims of mitochondrial cytopathy, 89 showed symptoms of the disease. This is a much higher ratio than would be expected if Mendelian segregation were at work.

Egger and Wilson conclude that mitochondrial cytopathy is transmitted by the DNA of the mitochondrion. In accounting for the fact that mitochondrial cytopathy was transmitted by the father in a few cases, they propose that occasionally a mitochondrial gene could be taken up into the nuclear genome (the full complement of genetic information in the nucleus). Once inserted into a chromosome the gene would be transmitted according to the rules of Mendelian inheritance.

If all the mother's mitochondrial DNA had the same defective nucleotide sequences and the gamete got copies of all the maternal mitochondrial genes, then the children of a woman suffering from the disease would all have the syndrome with the same degree of severity. In Egger and Wilson's observations, however, there was considerable variation in the severity of the symptoms, and some children escaped the disease completely. It could be that not all of the mother's mitochondria contain the abnormal nucleotide sequences. When the gamete is formed, copies of the mitochondrial genes are parceled out to the egg according to non-Mendelian rules of distribution. The number of copies of the aberrant genes could vary from egg to egg, and hence the severity of the disease would vary among offspring.

Quantum Evaporation

Albert Einstein's explanation of the photoelectric effect, in which an electron is ejected from the surface of a metal by light or some other form of electromagnetic radiation, was a decisive step in convincing many physicists that Max Planck's hypothetical quanta, or discrete packets of energy, are physically real. One of the most straightforward analogues of the photoelectric effect is the evaporation of atoms or molecules from the surface of a liquid. Under normal circumstances, however, such a large number of atoms or molecules are ejected from the evaporating surface that the presumed quantization of

the thermal energy causing the evaporation is smoothed out. Now three experimental physicists at the University of Exeter, M. J. Baird, F. R. Hope and A. F. G. Wyatt, have designed a sensitive apparatus that enables them to observe quantum evaporation under extreme physical conditions. In quantum evaporation a single atom is ejected from the surface of a liquid by a single, quantized thermal excitation of the liquid. The work is reported in *Nature*.

The investigators point out that ordinary evaporation remains far too complex to be explained from the first principles of the quantum theory. The difficulty is to understand how atoms acquire the energy needed to overcome the forces binding them to the surface of the liquid, an energy that is often far in excess of the average kinetic energy of an atom in thermal equilibrium with the rest of the atoms in the liquid. Hence in order to demonstrate quantum evaporation the complex interactions that give rise to ordinary evaporation had to be eliminated; experimental conditions were imposed in which the mechanism of evaporation would match, as closely as possible, the mechanism proposed by Einstein to explain the photoemission of electrons.

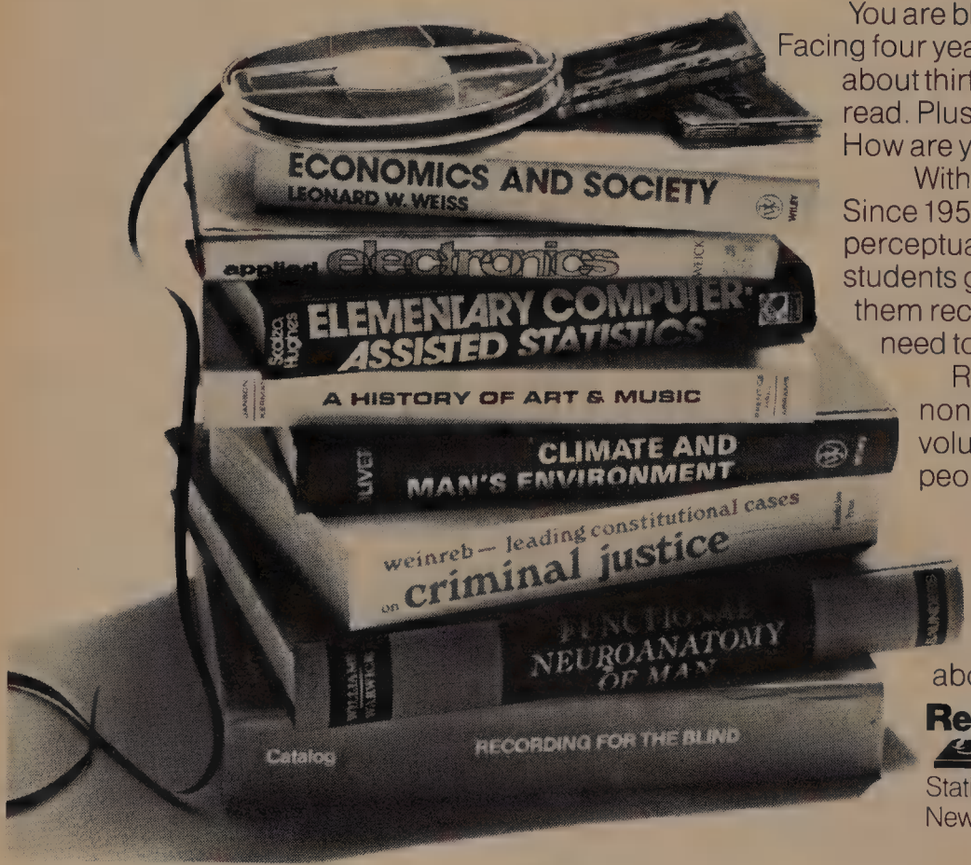
The quantum nature of the photoelectric effect is evident in the frequency threshold that must be crossed before any electrons are emitted. Furthermore, above the threshold frequency the kinetic energy of the emitted electrons varies directly with the frequency of the incident radiation. A change in the intensity, or amplitude, of the incident radiation, however, is inconsequential because an electron can gain energy only in the discrete amount carried by a single photon, or quantum of radiation. The energy content of a photon varies directly with its frequency, and so only photons of sufficiently high frequency can eject electrons from a metal. The kinetic energy of the emitted electron is equal to the difference between the energy transmitted to it by the incident photon and the binding energy the electron must overcome in order to escape from the metal.

The quantum-mechanical analogue of the photon for thermal excitations is the phonon. As early as 1912 Peter J. W. Debye (and almost simultaneously Max Born and Theodor von Kármán) recognized that phonons, very-high-frequency pressure waves, could transmit thermal energy through a liquid or a solid. Although from a classical (non-quantum-mechanical) point of view phonons are merely distortions of the atomic medium, quantum mechanics implies that the phonon must also be understood as a particle that moves through the medium with a certain wavelength and carrying a discrete amount of energy.

The major difficulty overcome by the Exeter workers was in demonstrating that the evaporation of an atom can be caused by a single incident phonon and not by some more complex interaction. In many liquids the mean free path of a phonon, that is, the average length of the path before the phonon is scattered or attenuated, decreases sharply with an increase in the phonon's energy. If a phonon with sufficient energy to evaporate an atom were experimentally introduced into such a liquid, its mean free path would be so short that by the time it reached the surface of the liquid its integrity could no longer be guaranteed. If evaporation were to take place, a mechanism for the evaporation analogous to that for the photoelectric effect could not be inferred. In a liquid consisting of the isotope helium 4, however, the mean free path of a phonon becomes quite large for a range of phonon frequencies above a critical value; the critical frequency is greater than the frequency required to eject atoms from the liquid-helium surface.

Baird, Hope and Wyatt submerged a small heater in a bath of liquid helium 4 cooled to within a tenth of a degree Celsius of absolute zero. The heater was

Close your eyes. Now have someone read this to you.



You are blind. A student. Facing four years of college. With about thirty-two textbooks to read. Plus fifty supplemental texts. How are you going to manage?

With Recording for the Blind. Since 1951, we've helped over 53,000 blind, perceptually and physically handicapped students get through school. By sending them recordings of the books they need to read. Free.

Recording for the Blind is non-profit, and supported by volunteers and contributions from people like you who can imagine what it's like to be blind.

Your tax-deductible donation will help our students meet their educational goals. We'd all be grateful.

If you want to know more about us, write:

Recording for the Blind, Inc.
an educational lifeline.
Station E, 215 East 58th Street
New York, New York 10022, (212) 751-0860.

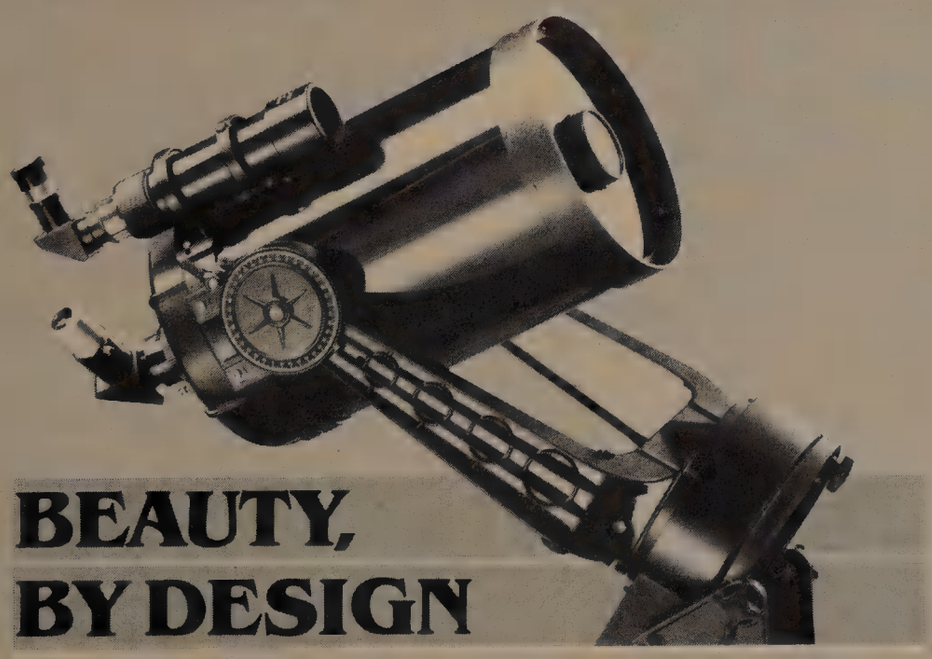
connected by a rigid bar to a bolometer above the surface of the helium that could detect helium atoms emitted from the surface. The bar was fixed to a carriage, and so the entire apparatus could be moved up or down with respect to the surface of the liquid helium. The distance between the heater and the bolometer was measured, and the time between the emission of a pulse of energy from the heater and the first detection of helium atoms at the bolometer was recorded for numerous vertical positions of the heater.

The pulse of energy at the heater introduces phonons of many different frequencies into the liquid helium. The velocity of the phonons through the liquid varies inversely with their wavelength, so that the low-frequency phonons arrive at the surface of the liquid first. On the other hand, for phonons that are energetic enough to evaporate atoms from the surface, the higher the frequency of the phonon, the greater the velocity of the emitted atom and the smaller the time of flight of the atom from the surface of the liquid to the bolometer. Hence for a given depth of the heater in the liquid there is a characteristic interval between the heating of the liquid and the arrival of helium atoms at the bolometer, provided the phonon excitations travel unimpeded through the liquid. The measured times for many different depths agree well with the predicted values, and so the experimenters conclude that the energy required for evaporation is transmitted by individual phonons. They write: "Liquid helium 4, therefore, is the first liquid in which we can understand evaporation at the atomic level."

Dynamic Proteins

Every textbook of molecular biology includes dozens of drawings of molecules. In some of them balls representing atoms are joined rigidly by sticks representing bond lengths and angles; in others rigid three-dimensional forms represent the volume of influence of each atom's electrons. In a way the drawings are misleading: they show only the average form of the molecule over a long period of time. In actuality the molecules of importance in molecular biology tend to be fluid. For example, globular proteins, which consist of compact, serpentine chains of amino acids, are flexible: the bonds that link the succession of amino acids can rotate, so that over times on the order of 10^{-13} second "the motion of individual groups" in the protein is "similar to that of small molecules in a liquid. The groups jiggle about as a result of frequent collisions with the surrounding 'cage' atoms" in the liquid. Over longer periods "larger regions in the protein molecule" make "slower elastic movements."

The above quotations are from a discussion of the fluid motions of globular proteins by J. Andrew McCammon of the University of Houston and Martin Karplus of Harvard University. In an article in *Accounts of Chemical Research* McCammon and Karplus treat the subject as a new theoretical science. The first step, they write, is to devise a mathematical description of the potential energy of a protein as a function of the positions of its atoms. The description includes terms depending on bond lengths and bond angles. Additional terms represent interactions between atoms in the protein that are not bonded to each other. Such interactions are "important in globular proteins because of the dense packing of atoms in these molecules." Indeed, the "close-packed nature of globular proteins" makes their short-term properties similar to those of a "simple liquid" in which "molecules are composed of hard sphere atoms; that is, the structure is one that optimizes the packing of such molecules, and the molecular motion reflects the frequency of collisions between them." The completed description of the potential energy of the protein readily yields a



BEAUTY, BY DESIGN

You are looking at the newest design of the world's most popular modern telescope — the Celestron Super C8. A telescope that many people, of similar mind to you, already own.

It doesn't look like a traditional telescope because it isn't a traditional telescope. The Super C8 uses advanced "folded optics" of Schmidt-Cassegrain design. Very much like the optical designs used by today's large observatory telescopes. Like them, it employs computer-designed, laser tested optics.

The Super C8 uses an 8" primary lens (that captures over 700 times more light than the human eye) and "folds" the image, to gain 2,000mm of focal length in a tube 16 inches in length.

It is lightweight, portable, compact and an extraordinary value. And it's beautifully adaptable for astrophotography and terrestrial viewing. The Celestron Super C8 can go from showing you and your guests distant nebulae, galaxies and star clusters, to the rings of Saturn, the moons of Jupiter or the wonders in this world.

Each Celestron Super C8 comes complete with accessory optics, high transmission coatings, easy-to-use fingertip adjustments, an ultra-precise Byers worm gear clock drive (to correct for the earth's rotation; ■ necessity in an instrument this powerful). And each comes with Celestron Certified Optics and a 25 year limited warranty.

If you would like to know more about Celestron, our new Super C8, or our complete line of binoculars, spotting scopes and telephotos, call the toll-free number listed below for information and your nearest dealer.

CELESTRON®
PRECISION OPTICS

For Dealer Locations, Catalogs, Information, Technical Data Sheets

CALL FREE
1-800-421-1526
(Mon-Fri; 8-4:00 Pacific Time: Continental USA only)

CELESTRON INTERNATIONAL, P.O. Box 3578 SA10, 2835 Columbia St., Torrance, CA 90503 (213) 328-9560

mathematical description of the forces acting on each atom, and then, by Newtonian equations, the description of the forces yields a description of how the molecule moves over time.

Slower motions in the protein require different methods. In one method the motion of "a group of side-chain atoms in a protein" is taken to be analogous to the diffusion of a molecule through a liquid or a solid. "It is therefore a temperature-dependent quantity and typically becomes softer, or more permissive, as the mobility of neighboring groups increases." Again the interactions between atoms that are not bonded to each other prove to be important. "These interactions produce randomly varying forces that act to speed or slow the motion of the group in a given direction." The motion itself entails a certain degree of friction between the group and its surroundings, and that too must be taken into account.

The dynamics of certain parts of certain proteins have already been modeled in detail. Bovine pancreatic trypsin inhibitor (BPTI) is an example. The atoms in BPTI turn out to fluctuate about their average positions by distances comparable to the radii of the atoms. The fluctuations "increase markedly near the protein surface." The 21st amino acid in the chain of BPTI is tyrosine, which is notable for a ring of carbon atoms projecting outward from the backbone of the chain. The ring turns out to twist about its bond to the backbone; "the motion is analogous to the rotation of a benzene [ring] about an axis passing through two atoms at opposite vertices of the ring." In BPTI the time-averaged amplitude of the twisting (about 12 degrees) is limited by "hindrance between atoms in the ring and those in the cage surrounding the ring," but "large-scale distortions of the protein produce small changes in the overall shape of the cage, with a concomitant alteration of the ring orientation." Sometimes the ring flips: it twists 180 degrees. The theoretical studies suggest that the flip requires only .5 to 1 picosecond and that it occurs "not as the result of a particularly energetic collision with some cage atom but as the result of a transient decrease in frequency and intensity of collisions that would drive the ring away" from having the energy a flip requires.

McCammon and Karplus suggest that "the dynamical nature of protein molecules is essential to much of their biological activity. Indeed, it is reasonable to suppose that the internal flexibility of proteins and some of their dynamical characteristics have been shaped by evolution to allow efficient expression of protein function." It is known that the protein myoglobin, which binds oxygen molecules in muscle cells, has "no channel for oxygen entry." The entry "is dependent on transient structural fluctuations." On larger scales the hinged motion of parts of a protein may modulate the rate at which the protein acts, say, as an enzyme, that is, as a catalyst. The theoretical study of protein dynamics should make it possible to predict such rates "before the end of this century," and the predictive ability "will ultimately be of practical use in connection with genetic engineering and industrial enzyme technology."

Star Grazers

Two workers at the Lunar and Planetary Institute in Houston studying the images from the 1976 Viking mission to Mars noticed something funny about the craters that pit the Martian surface. Too many of the craters were elongated rather than circular, and the direction of the major axis of the elongated craters seemed to vary systematically with their age. A catalogue of 176 such craters confirmed this impression and led Peter H. Schultz and Anne B. Lutz-Garihan to some startling conclusions about the bodies that produced the craters and the reason for the variation in the axis of the craters. They present them in a recent issue of *Journal of Geophysical Research.*

Most impact craters on the moon, Mars and Mercury are relatively circular and are surrounded by a symmetrical pattern of ejecta. An elongated crater with a butterfly-wing pattern of ejecta perpendicular to the long axis of the crater is produced only by a grazing impact. Laboratory experiments have shown that the angle of impact must be less than five degrees to make craters with both these identifying characteristics. For objects in orbit around the sun, such as asteroids or comets, the most probable angle of impact with another body is 45 degrees. The probability of impact at angles less than five degrees is .7 percent. On the maria (lava plains) of the moon there is only one grazing-impact crater larger than three kilometers, a frequency consistent with the theoretical probability of grazing collisions between heliocentric objects and the moon. On the lava plains of Mars, however, about 5 percent of the craters are grazers, a frequency 10 times higher than collisions with heliocentric bodies should have produced.

When the 176 Martian grazers were grouped according to their age (determined on the basis of the degree to which they had been eroded), it became apparent that the major axes of craters of roughly the same age often fall on the same great circle and that the great circles traced by the axes of older craters are increasingly inclined toward the Martian equator. In other words, the axes of newer craters are roughly east-west with respect to the Martian geographic poles and the axes of older craters are increasingly north-south with respect to the poles.

After considering several possible interpretations of the evidence Schultz and Lutz-Garihan conclude that the grazers were produced not by heliocentric bodies but by areocentric ones: a large family of Martian satellites whose orbits decayed over time. The consistency of the orientation of the axes of impact indicates that at any given time the satellites had orbits with nearly the same inclination. This suggests that a swarm of smaller satellites may have been created by a process such as the incomplete accretion of one large satellite, at least 225 kilometers in diameter. The tiny Martian moons Phobos (which will collide with Mars in about 10 million years) and Deimos may be the last members of this large family.

Schultz and Lutz-Garihan also consider two possible explanations for the variation of the axis of cratering with time. They conclude that the axis changed not because the inclination of the orbits of the satellites changed but because the Martian crust moved. Any of several processes causing substantial changes in the moment of inertia of the planet (the distribution of mass with respect to the spin axis) could have caused the crust to migrate. For example, the orientation of the long axes of the grazing craters shifted from north-south to east-west at about the same time as the first major epoch of Martian volcanism began. The volcanic and tectonic processes that formed a large uncompensated mass such as the Tharsis province on the surface of the planet and the upwelling of dense mantle material that may have accompanied the formation of the mass would have significantly changed the moment of inertia. Centrifugal force would then have caused the mass to migrate toward the equator (outward) until the greatest mass (the principal moment of inertia) coincided with the equator and the smallest mass (the minimum moment of inertia) coincided with the planet's pole or spin axis. The orientation of the axes of the Martian grazers provides a clear record of the redistribution of mass with respect to the spin axis caused by the migration of the crust.

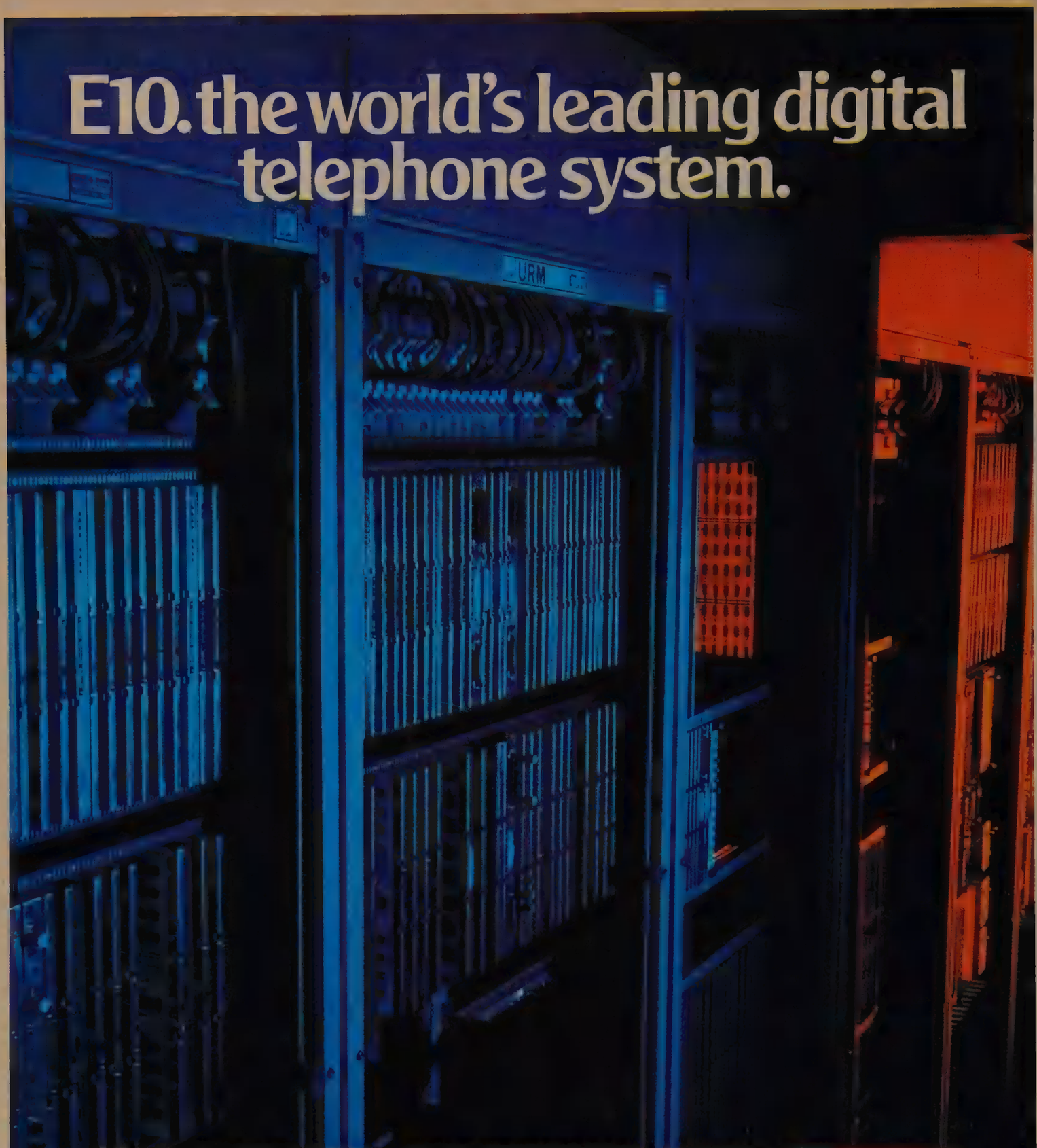
It can be shown on this basis that the points on the Martian surface that are now the geographic poles were once at much lower latitudes. In general the variation of the axis of cratering records the wandering of the Martian geographic poles much as the alignment of magnetic grains in the rocks of the earth's crust records the wandering of the earth's magnetic poles. The Martian grazers might thus serve as the basis for reconstructions of the past geography of Mars just as remanent magnetism serves as the basis for reconstructions of the past distribution of the continents of the earth.

French Communications: Exporting Technical Expertise

This special report on French communications was written by Joel Stratte-McClure following visits to public and private research facilities, government agencies and industrial companies throughout France.



E10. the world's leading digital telephone system.



E 10, CIT ALCATEL's digital time-division switching system, has made a major contribution to the development of telecommunications worldwide.

Today, one out of three time-division lines is installed by CIT ALCATEL. E 10 has thus become the first digital switching system in the world with 11 million lines ordered in 33 countries, of which 5 million are already in service.

CIT ALCATEL, the first company in the world to introduce digital time-division switching, has devoted its 10-years of experience to developing a modern and proven system which

- open to the telematics applications (videotex, multiservice, mobile radiotelephony);
- capable of ensuring a completely digital connection throughout the line as the rate subscribers are digitalized;
- adaptable to the specifications of the future integrated services digital network (ISDN).

CIT Alcatel

CIT ALCATEL - 33, rue Emeriau - 75015 Paris - France - Tél. : (1) 571.10.10 - Télex : 250.927
TSS ALCATEL, INC. - 11410 Isaac Newton Sq. - Reston - Virginia 22090 - U.S.A.

CREATING A DOMESTIC INFRASTRUCTURE

One of the cornerstones of the French government's ambitious attempt to become a global force in high technology (see "French Telecommunications Digital Technology and the Télématique Program" and "French Technology: Preparing for the 21st Century," in the September 1980 and November 1982 issues of this magazine, respectively) is to enhance the lead it has established in numerous sectors of military and civilian communications. During the past decade France has developed and created a number of advanced products and systems, introduced a wide range of domestic communications services, launched a variety of long-term research and development (R&D) projects, contributed to international standards and penetrated the international marketplace.

The transition in French communications began in 1975, when the modernization of the domestic telephone network became a priority of the country's seventh five-year plan. $30 billion was initially allocated to improve the internal telecommunications network and increase the number of subscriber lines from 6.2 million in 1974 to over 20 million in 1983.

The technical emphasis was on the development and introduction of electronic time division multiplex (tdm) switching (the world's first electronic tdm switching center was operating in France in 1970), digital pulse code modulation (pcm) transmission, a packet-switched data network and a multiservice domestic satellite system. Government and industrial R&D concentrated on microelectronics, optoelectronics, intelligent terminals and networks, optical and satellite communication systems, computer and microprocessor assistance to network function and new communication concepts.

In the late 1970's, the French launched the *télématique* (a word formed by combining *télécommunications* and *informatique*, which means data processing) program. *Télématique* integrates numerous compatible products and services into a harmonious network permitting data and other forms of information to be transferred speedily using pcm transmission and tdm switching techniques. Today the emphasis in French communications is on utilizing the country's semi-digital network infrastructure to create easy-to-use low-cost *télématique* services for companies, community organizations and individuals. *Télématique* transforms the public switch telephone network (pstn) into a multifunctional tool that answers a variety of communication requirements.

In 1986, it will be possible for French subscribers to have an all-digital dialing network with a 64 kilobits per second (kbit/s) transmission rate providing access to a fuller range of new services. Digitization of the network to the subscriber level will produce a digital synergy merging terminations, private branch exchanges (pbxs), cellular radio and satellite services with new forms of transmission—primarily satellite, fiber optics and packet switching. The result will be the concurrent storage, processing and transmission of text, data and voice on a single carrier.

In addition, the government recently decided to complement this integrated telephone and data network (itdn) with a nationwide fiber-optical videocommunications network—which will pave the way for the implementation of a wide-band integrated services digital network (isdn). The French are gambling on its also enabling them to leapfrog their competitors on the technological front and further exploit export markets.

PTT: PROVIDING THE DIRECTION

The French communications effort is controlled by the PTT—the Ministry of Postes, Télécommunications and Télédiffusion—which owns, regulates and operates the national telecommunications network. The PTT directs and executes telecommunications, television and postal policy and is responsible for the pace of technical change, the type of equipment permitted on the market and the services available.

The Direction Générale des Télécommunications (DGT) is the telecommunications arm of the PTT and implements telecommunications policy. Within the DGT are the Direction of Industrial and International Affairs (DAII), the National Telecommunications Research Center (CNET) and other affiliates responsible for overseeing the effort.

While some PTT's are moving towards deregulation, there is no intention of further liberalizing telecommunications policy in France. In 1982, the government—continuing a process begun in the mid-1970's, when some foreign firms were forced to sell their telecommunications operations in France to local companies—nationalized major telecommunications, data processing and electronics firms.

"Nationalization will permit more collaboration, promote efficiency and provide a unified effort," says Louis Mexandeau, Minister of the PTT. "It helps France avoid duplication of R&D and speeds up development. In addition, industry will remain tightly regulated because the alternative leads to an absence of any global planning and guidance—often resulting in higher charges and deteriorating service."

Although the current French economic austerity program has led to budget reductions throughout the public sector, the PTT remains a major innovative force in the country's development.

"Our charter provides that the budget will remain at 27 billion French Francs (FF)* for the next four years," says Mexandeau. "The primary expenditures will be development of the regular telephone network, *télématique* services and wide-band cable technology."

Jacques Dondoux, the Director General of Telecommunications, contends "France has introduced *télématique* products and we intend to make *télématique* profitable, establish a competitive fiber optic industry, launch our domestic satellite and increase telephone traffic."

THE EMPHASIS ON ELECTRONICS

The electronics industry, with 1,700 companies and a workforce of 490,000, is the keystone of the French industrial program. As part of a government plan, 140 billion FF is slated to be invested in electronics by the state and industry during the next five years, with 80 billion FF earmarked for telecommunications and professional electronics.

The key goal is to reduce the negative trade balance in electronics products, which rose from 6 billion FF in 1981 to 12 billion FF in 1982. Another aspect of the plan is to create a strong domestic industry and boost the French presence in export markets. France represents about 5 percent of the world electronics market, but 52 percent of the French electronics and telecommunication markets is controlled by foreign companies—primarily from the U.S. (22 percent), Japan (7 percent), West Germany (7 percent), and Holland (6 percent). Excluding their domestic market, the French have only 1.4 percent of the world electronics market—and only 0.3 percent of the U.S. electronics market and less than 2 percent of the U.S. telecommunications market. The electronics effort is meant to alter this situation.

"The solid structure of French telecommunications will bring a dynamism to other electronics sectors," predicts Mexandeau.

Critics contend that the French elec-

*At the time of writing, 1 U.S. dollar equals approximately 8 FF.

ADVERTISEMENT

tronics plan, much of which is under the supervision of the PTT, is too vast and financially unrealistic. But most observers agree that France has a technical lead in telecommunications that can be exploited.

THE INDUSTRIAL STRUCTURE

The French want to have at least two domestic suppliers in most areas of telecommunications rather than create a nationalized "Telephone de France." Although some foreign companies have participated in the development of French telecommunications, most of the effort is allocated to domestic companies.

Five companies, nationalized or partly owned by the French state as part of President François Mitterrand's Socialist program following his election in May 1981, are the prime industrial movers behind the communications effort: Bull, Compagnie Générale de Constructions Téléphoniques (CGCT), CIT Alcatel, Matra and Thomson-CSF.

CIT Alcatel is a subsidiary of the Compagnie Générale d'Electricité, (CGE), the fifth largest company in France, and has always conducted a strategic concertation with the government.

"The nationalization has had a favorable influence because it provides the long-term outlook needed in the telecommunications industry," says Pierre Chavance, CIT Alcatel's director general.

Today CIT Alcatel is a leading manufacturer of tdm switching systems (30 percent of the world market), 30-channel pcm equipment, submarine telephone systems and electronic pbxs. Its affiliate Alcatel Electronique is a specialist in mail handling, peripheral telecommunications equipment, radiocommunications and electronic telex terminals. SESA, in which CIT Alcatel has a 55 percent stake, is the leading French company in the area of packet switching networks, and its affiliate Câbles de Lyon is France's largest manufacturer of telecommunications cables.

Thomson-CSF—part of the Thomson group and a leading exporter of detection systems, microwave equipment, satellite earth stations and radio/TV transmitters—has about 40 percent of the French tdm public telephone switching market and 21 percent of the private telephone market. Thomson has recently restructured its telecommunications activities, placing a priority on forming more international agreements and penetrating the U.S. market.

"We have synthesized the wide range of our activities in communications to provide customers with complete turnkey systems in numerous areas," says Jacques Darmon, the head of Thomson-CSF's communications branch.

The Thomson-CSF communications branch includes public telecommunications, radio communications, business communications, knowledge information processing systems and software.

Matra—a diversified defense, electronics and communications group—is consolidating its telecommunications industrial base for increased development of pbxs and terminals. Participation by the French state in Matra has not altered the company's ambitious diversification philosophy, which has led to numerous international joint ventures.

"The French government has not replaced our management as it has with other newly nationalized companies," says Maurice Remy, director of Matra's telecommunications branch.

CGCT, a former subsidiary of ITT, previously manufactured space division electronic switching systems. After two years of government debate, however, it has been given a license to manufacture the Thomson-CSF range of digital switching systems. CGCT also specializes in private telephone sets and systems and manufactures a range of pbxs, telephone and *télématique* terminals, teletex equipment, decoders and local area networks.

"We needed both government money and the decision to manufacture a digital switching system in order to survive," says CGCT president Pierre Lestrade.

Bull (formerly Cii-Honeywell Bull) is the second largest European vendor of data processing equipment. The French government has signed a comprehensive agreement with Bull to develop a complete range of data processing products, including software, components, mini- and microcomputers, office automation networks and medium-sized and large computer systems. Honeywell retains just under 20 percent of Bull and has renewed an existing technology and marketing agreement for ten years.

"Our task is to integrate our wide scope of activities and participate in the rapid developments taking place in French communications," says Claude Boulle, director of Bull's networks and communications group.

The group G3S includes CSEE (Compagnie de Signaux et d'Enterprises Electriques), SAGEM (Société d'Applications Générales d'Electricité et de Mécanique) and SAT (Sociéte Anonyme de Télécommunications). Forty-three percent of G3S activity is in telecommunications and data processing areas, including local and long-distance cables, microwave radio links, transmission equipment, teleprinters and teletex, facsimile transmission equipment, telephone and telegraph electronic switching, telemetry, telephone networks, cable TV, automatic postal sorting equipment and disk and bubble memories.

"Although there was initially concern that the government would primarily support nationalized companies, we are still getting our fair share of work from the PTT," says Francois Le Menestrel, chief executive officer of SAT. "We've maintained our market share in cables, transmission and microwave activites and export 30 percent of our turnover."

Jeumont-Schneider, part of the Empain Schneider group, produces components and software, markets a line of electronic pbxs and manufactures a rural tdm switching system.

TRT (Télécommunications Radioélectriques et Téléphoniques), an affiliate of Philips, is active in microwave radio systems, radio communications, data communications and switching, telephone transmission, *télématique* terminals and integrated digital systems.

INDUSTRIAL DEVELOPMENTS

Though the list of French communications projects is extensive, it is worthwhile surveying some of the more important activities.

INTEGRATED CIRCUITS: French companies have licensed chip-making technology and basic chip design from abroad, creating joint ventures and technology cross-licensing agreements with American firms. Following an expensive five-year integrated circuit plan, France has reorganized its components industry around two industrial poles, Thomson-CSF Components and Matra-Harris Semiconducteurs.

In 1982, French manufacturers accounted for only 5 percent of the domestic microprocessor and MOS memory markets (45 percent if foreign groups are included) but the French government will spend 4.5 billion FF on the integrated circuit industry from 1983 to 1986 in an effort to equalize the trade deficit.

SOFTWARE: France, Europe's largest supplier of software, developed the ADA language, powerful software for tdm switching (which requires about 1.2 million instructions) and software for *télématique* and videocommunication applications.

"France is effectively disseminating

the use of computers that will encourage more software advances," says Michel Jalabert, vice president of corporate development at information engineering services company Cap Gemini Sogeti.

Cap Gemini Sogeti and Steria have provided most of the software for the country's *télématique* videotex experiments; France Câbles & Radio subsidiary Télésystems plays an active role in data processing and produces software for electronic mail, network engineering, applications software and videotex; Sligos has developed software packages for the smart card; and Bull is expanding its range of sophisticated software for its data processing products and systems.

DIGITAL SWITCHING: The last electromechanical telephone exchanges in France were ordered in 1981, and the PTT continues to invest heavily in semi-digital and digital exchanges for cost-effective expansion (tdm is 30 percent less expensive than space division switching) and modernization of the telephone network. In 1983, more than 90 percent of the DGT's orders for switching equipment will be tdm, and more than 3 million digital lines are already in service.

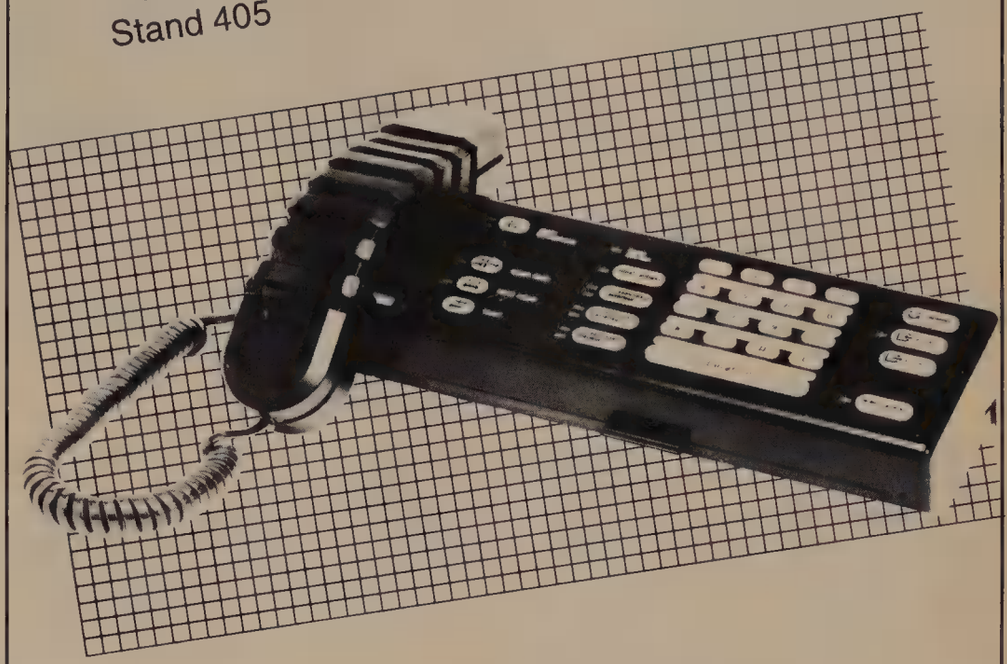
The results of digital switching are numerous—better service, simplified engineering because of its modular structure, sophisticated administration and maintenance, and enhanced features—and the PTT is now ordering second-generation systems from CIT Alcatel (E10 B local exchanges and E 12 transit centers) and Thomson-CSF (MT 20 transit centers and MT 25 local exchanges).

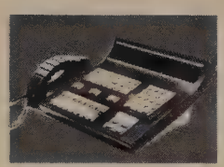
CIT Alcatel, whose E 10 switching system has been ordered by 34 countries, has a technological and commercial lead on Thomson-CSF, whose MT range of tdm exchanges has been ordered by 16 countries. CIT Alcatel's E 10 all-electronic telephone exchanges won 50 percent of the orders from the PTT last year against 35 percent for Thomson-CSF and 15 percent for CGCT. But both companies must now concentrate on the international marketplace, where the real marketing battle is being waged.

French industry has also developed modular, medium-capacity digital systems for telephone and *télématique* applications, including the E 10 S from CIT Alcatel and the MT 35 from Thomson-CSF. Jeumont-Schneider's JISCOS electronic public switching exchange has been designed for low- and medium-density population areas, and SAT manufactures the TC 300 digital public switching system with a capacity of over 8,000 subscribers for rural

OFFICE AUTOMATION CALLING... ON JEUMONT-SCHNEIDER TELEPHONE LINES

TELECOM 83
Hall 4
Stand 405








1975: A new impulse is given to private telephony, JEUMONT-SCHNEIDER launches its first time-division PABX. **1975-1982:** In France and throughout the world, thousands of companies of all sizes buy and install JISTEL and TLC systems representing nearly 1.5 million lines. **1983:** JEUMONT-SCHNEIDER consolidates its lead by launching the JISTEL 30 X, an advanced terminal system, and the JISTEL 500, a multi-function, fully-digital PABX which can handle voice, data and texts simultaneously.
To prepare your Company for the future, 700 engineers and technicians are working on tomorrow's communications and, thanks to JEUMONT-SCHNEIDER's network of agents, the most extensive that exists, you will always have competent servicing nearby. Consult us today. From a few to 8000 extensions, we have the answer to your specific needs.

OFFICE AUTOMATION CALLING... ON JEUMONT-SCHNEIDER LINES.

JEUMONT-SCHNEIDER

31-32, quai de Dion-Bouton 92811 Puteaux Tel. (33) 1 776.43.23 Telex 610 425 Melec

or medium-density zone telephone networks.

French companies are examining the evolution of digital switching equipment. "There will be a definition of new products based on applications such as animated images," says Jean-Claude Grynberg, head of the industrial switching group at the DGT. "Once manufactured, systems will then be tailored to individual needs."

"Industry is moving towards decentralized systems of autonomous subscriber connection units," adds Jacques Hauri, product planning manager of Thomson-CSF's public switching division. "There will be an integration of services with a more flexible system."

CABLE AND TRANSMISSION: Accompanying the introduction of tdm switching is the corresponding pcm digital transmission supporting environment. The French are converting the transmission system from analog to digital techniques that use hierarchical operating channels to transmit over coaxial cable, microwave radio relay and optical fibers. Digital transmission reduces the required amount of coding and multiplexing equipment, permits more circuits to be established on existing cables and facilitates transmission of new services.

The three primary cable manufacturers are CIT Alcatel, SAT, and Thomson-CSF subsidiary LTT (Lignes Télégraphiques et Téléphoniques). In addition, SAT, Thomson-CSF, and TRT compete for the microwave market; TRT and SAT are leaders in the modem market; and multiplexer and amplifier suppliers are CIT Alcatel, SAT and LTT. For several years, the DGT orders for transmission systems have been for digital equipment.

The major effort in the transmission field is to equip France with fiber-optical networks, including interurban links, wide-band links to subscribers, submarine cable and optical fibers for video telephone and video teleconferences. Most manufacturers provide cable and optical fiber telecommunications equipment, including long-haul analog and digital transmission systems for land links; cable for local networks and subscriber connection; equipment for *télématique* transmission, including the connection of professional terminals, facsimiles and data transmission equipment; and optoelectronic components.

PRIVATE TELEPHONY: Numerous French companies—CGCT, CIT Alcatel, Jeumont-Schneider, Matra, SAT and Thomson-CSF—are producing multifunctional pbxs with digital technology that save space and improve reliability. They include the basic components of a controller with sophisticated software, a switch matrix, a number of interface modules and transmission at 64 kbit/s. Sales of digital pbx equipment are expected to surpass analog in 1984—simultaneously creating a large market for vlsi chips, coder-decoders and subscriber line interface circuits (slics).

"There is a competitive, open market for pbx equipment in France, but the return is not great because the cost per line is much less than in the U.S.," says Jeumont-Schneider's commercial director Jacques T. Darmon. "We believe there is a great role in the office for the multiservice pbx as a local area network."

A number of companies are marketing the T 83, the new electronic telephone set available to French subscribers, as well as a variety of terminals for domestic use and export.

"The leadership for requirements and needs in private telephony comes from the U.S., but the market is worldwide," says François Dufaux, director of Thomson-CSF's private telephone division.

VIDEOTEX: Two French videotex systems are operating today. The Télétel interactive videotex uses the telephone line to transmit requested information, and the ANTIOPE (Numerical Acquisition and Televising of Images Organized in Pages of Writing) teletext broadcasts information repetitively in the blanking interval between frames

FRENCH COMMUNICATIONS:

SMALL COMPANIES ARE PROSPERING

While nationalized and large private companies in France handle a substantial amount of telecommunications activity, a number of smaller firms are also making considerable technical breakthroughs and creating an international commercial impact.

"We are trying to find the best synergy possible between the investment and research capabilities of nationalized firms and the innovative capacity of smaller companies," says the DGT's Jacques Dondoux. "Innovations range from products associated with digital and packet switching to the *télématique* arena and private telephony."

Some examples:

—Barphone, located in Saumur in the Loire Valley, has established a worldwide commercial network to market its telephone terminals and accessories, which include fully electronic key sets.

—BOGA, with manufacturing facilities in the Normandy town of Fécamp, specializes in equipment for the interface between subscribers and the automatic telephone exchange.



FIET

Numerous small French companies are active in telecommunications. This portable minitel terminal is manufactured by FIET.

—FIET (France Informatique Eléctronique & Télématique), with a plant at Chateaugiron in Brittany, manufactures a 26-pound portable terminal that houses a 5.5-inch pop-up black-and-white CRT screen, a full alphanumeric keyboard and a 40 character per line matrix printer.

—Tekélec has developed an X.25 software package for its TE 92 packet switching analyzer simulator that has become a model for diagnostic software for public and private networks. Its pcm and data transmission testing instruments are manufactured in Pessac, near Bordeaux.

—Visioptronic in Paris manufactures videophones and other products for video communications systems, including wide-band intercoms, cables and connectors.

—XCom in Grenoble is working on a variety of *télématique* products and applications, including computer translation. The company has developed a system for speech synthesis from text, using a program designed to calculate the intonations of sentences and phrases from their punctuation marks. The S-100 card, employing a system of diphones, is based on a CNET speech synthesis coding circuit and an 8085 microprocessor. The card could be used for distribution of messages from a data bank, creation of public service messages, man-machine dialogue and industrial processes.

The important role of small companies in telecommunications cannot be emphasized enough," says Dondoux. "They are the inspiration for the industry."

on the television channel.

One aspect of the Télétel videotex system is the electronic telephone directory, inaugurated in Britanny last February and progressively available in most other French cities. The electronic directory program could be considered the largest conversational data processing system in the world. Six hundred thousand minitel interactive terminals (minitel is the generic name for a range of stand-alone interactive videotex terminals) have been ordered by the PTT, and approximately 140,000 terminals will be in service by the end of 1983. By 1990, the PTT says, every Frenchman will be able to exchange the telephone book for a free minitel. This volume provides for relatively low-cost terminals compatible with other videotex services.

Cap Gemini Sogeti, Matra, TRT and Bull are developing one system, while Bull, CIT Alcatel, SESA and SINTRA Alcatel are producing a second. The extension on a national basis will be the responsibility of Cap Gemini Sogeti, SESA and CIT Alcatel. SESA and IBM are also working with the PTT on a computerized directory service for operators.

"The directory, which will enable the user to obtain telephone numbers throughout France, is not a goal in itself but a means of communications opening the way for other *télématique* services," says Jean-Paul Maury, head of the project at the DGT.

Télétel, which combines the electronic directory with other interactive services, was tested in 2,241 households near Paris and included timetables, news, teleshopping, stock quotations and information from other data banks provided by 190 service suppliers. A study of users indicated that the average consultation lasted 14.5 minutes, accessed 3.6 data bases and involved transmission of about 47 television frames of information.

Some French firms are using portable videotex terminals to provide sales representatives wih instant two-way access to computerized records. The towns of Grenoble and Nantes are running municipal videotex systems, and the Crédit Mutuel de Bretagne is using the minitel for home banking services.

THE SMART CARD: The smart card is an intelligent plastic credit card containing an imbedded integrated circuit capable of storing information and conducting numerous transactions without costly interfaces or communication with a central computer. It can be employed as an electronic checkbook for point-of-sale purchases, a telepayment card to conduct financial services over a videotex system, a high-security au-



CIT Alcatel/Thomson-CSF

The latest generations of digital switching exchanges manufactured by CIT Alcatel and Thomson-CSF are the E 10 (left) and the MT 20, photographed respectively in Lebanon and Jordan.

FRENCH COMMUNICATIONS:

SUBMARINE CABLE

Although satellites are obviously an important factor in the future of telecommunications transmission, numerous countries will continue installing submarine cable systems during the 1980's. France helped spearhead underwater links in 1957 using coaxial cable, and over 40,000 kilometers of French-manufactured submarine links are in service. Manufacturers are proposing systems with capacities ranging from 500 to 5,000 channels.

Submarine transmission is ideal for conveying traffic requiring large circuit capacities over short and medium-distance point-to-point links. Existing high-performance capacity submarine cable systems have improved transmission quality and reliability and reduced maintenance. And while capacity has increased, the cost of submarine circuits has decreased. The price per kilometer of circuit for a transatlantic telephone channel has declined from $140 in the 1950's to $9 today.

One important project currently taking advantage of these improvements is the largest analog submarine cable ever installed. The SEA-ME-WE (Southeast Asia-Middle East-Western Europe) cable will be 14,000 kilometers in length and link France with Singapore when it is completed in 1985.

The introduction of fiber-optical submarine transmission will meet increased intercontinental traffic demands and play an important role in establishing networks capable of various teleprocessing applications. Numerous government and commercial studies are under way in France to extend digital transmission on optical fibers. An optical fiber cable has been laid on the French Riviera, and a submarine fiber-optic link between Corsica and France with a flow rate of 2×140 Mbit/s is scheduled to come into service in 1985.

The French have played a major role in the evolution of submarine cable technology—aimed at improving transmission quality, reducing costs and increasing traffic-handling capacity. Their effort has resulted in substantial improvements in the manufacture of cable components—including the cable itself, repeaters, equalizers and cable-laying equipment.

R&D on submarine cable is geared towards decreasing attenuation by improving the quality of the structures and materials used for conductors. Repeaters, the submerged units used to amplify signals, have increased gains and passbands, broadband transistors have been developed, and automatic gain regulation systems have been introduced. Improved equalizers permit the remote control of characteristic frequencies from terminal stations and allow reliable remote monitoring and fault locating.

Submarcom, a joint venture of CIT Alcatel and Câbles de Lyons, bid for a role in the $300 million contract for the first transatlantic cable (TAT-8) using optical fibers. The cable, slated for deployment in early 1988, will handle 40,000 calls simultaneously and double the amount of voice and data traffic that can now be transmitted between Europe and North America. Other French companies, such as PTT affiliate France Câbles et Radio, provide cable-laying ships and equipment.

thorization and access card, a portable file card, a prepaid card for pay telephones, or a billing method for computer time sharing, pay television programs, or other electronic services.

The French launched three ongoing large-scale field trials with 120,000 cards and 650 point-of-sale terminals in three cities—Lyon, Blois, and Caen—as part of a 50 million FF test being financed by a syndicate of French banks. In addition, 300 users employed the smart card for teleidentification and telepayment during the Télétel tests.

Three French companies—Bull, Flonic Schlumberger and Philips Data Systems—are developing the card in collaboration with in-house or external software firms. Sligos, a systems house, has been contracted to operate the central clearing systems between banks. Although each firm has a different technical approach, an international association (INTAMIC) has been formed to monitor common standards for the smart card.

In addition, the U.S. Department of Defense is testing the smart card, with Army personnel using it to gain admittance to hospital and commissary facilities, while the Department of Agriculture is considering a test in which the card will replace food stamps.

The use of credit cards for electronic funds transfer and point-of-sale purchases is also expanding in France. Magnetic cards and multiservice terminals designed by Electronique Serge Dassault are being used in a 30-month trial in Saint-Etienne for instantaneous point-of-sale transactions. The system was designed by SG2 (Société Générale de Service et de Gestion) and the voice-over data equipment was produced by SECRE. An off-line test with magnetic cards is underway in Aix en Provence, using terminals manufactured by Electronique Serge Dassault and CSEE.

"The future card will combine the smart card and magnetic strip and have multiple functions with more than one microprocessor," predicts François Salle, group vice president for research and technology at Bull.

The PTT, which to date has ordered 450,000 smart cards, has also awarded development contracts for 200 public telephone booths to accept the smart card and bank credit cards—for billing through the user's post office or bank account. Prototypes of the booths were put into service last June. An additional 2,000 telephones using cards will be installed in 1984, and 10,000 are scheduled for introduction in 1985. There are currently over 150,000 public telephones in France; use of cards will prevent fraud, enable prepaid calls and reduce the cost of collecting 16,500 tons of coins.

"The expanding growth of applications will result in a reasonable price for the smart card," says Jean-Louis Marchand at the DGT's smart card delegation. "It should decline from 65 FF today to about 30 FF in 1986."

TELETEX: Teletex combines text typing and processing with directory-

FRENCH COMMUNICATIONS:

THE WORLD'S LARGEST X.25 PACKET-SWITCHED NETWORK

The French public packet-switched data network has substantially promoted the use of distributed data processing by providing access to computers from remote terminals—with tariffs independent of distance. The Transpac network, which is connected through the international transit node to the European Economic Community network (Euronet) and data networks in 20 other countries, has transparent features allowing users anywhere to be connected to the data storage and processing capabilities of large machines.

Almost any remote processing need can be met by Transpac. Its uses cover inquiries, updating files, time sharing and on-line data transaction processing. It can also handle off-line data collection with deferred transmission. Consequently, Transpac is an important factor in the expansion of videotex, teletex, high-speed facsimile, point-of-sale transactions and electronic mail. The videotex national network, for example, is based on Transpac and uses distributed gateways throughout France. Transpac is also the link between host computers in the electronic directory system and is used as the transmission vehicle for the point-of-sale tests in Saint-Etienne.



The Transpac packet-switched data network, with its operational center in Rennes, provides access to computers from remote terminals with tariffs independent of distance.

Transpac, providing the circuit between a customer's terminal equipment and the network, is articulated around data concentrating and switching processors. Information is submitted to the transmission network in the form of "packets," with network nodes forwarding the packets as they are received. The switching nodes are interconnected by a highly meshed network of fast data channels with at least two 72 kbit/s links between each pair of switches. Operation of the network is supervised by a national management center.

Transpac entered service at the end of 1978 with twelve switching centers and a capacity of 1,500 subscribers. Today there are over 11,000 direct subscribers and 22 switching centers in 20 cities, and transmission potential has been increased from 50 bit/s to 48 kbit/s. Eleven hundred additional users access the system through the pstn and telex networks, from 300 bit/s to 1200 bit/s.

A new type of switch, SESA's DPS 25, has been introduced to accomodate growth in network connections and traffic. It permits the installation of small satellite switches, a gradual increase in the number of main nodes and specialization of certain switches as transit nodes.

The Transpac network involves the participation of a number of French companies. Network designer SESA, TRT and Bull subsidiary SEMS are responsible for the switching system, SAT supplies the multiplexers, and CIT Alcatel, Thomson-CSF, SAT and TRT provide various categories of modem.

Two-thirds of direct connections to Transpac employ the X.25 protocol (the standard software interface for access), making it the world's largest X.25 network. Transpac transmits about 40 billion bits and treats 100,000 calls a day—and during peak hours handles 5,000 simultaneous communications. Transpac, managed by the PTT, will invest 300 million FF in 1983 to increase the capacity of existing concentrators and integrate new inter-nodes. The PTT estimates that with investments of 2 billion FF between 1983 and 1990, there will be 100,000 subscribers in 1990.

FRENCH COMMUNICATIONS: THE EVOLUTIONARY CAPABILITY OF THE E 10 DIGITAL SWITCHING SYSTEM

CIT Alcatel has been active in the communications business for over 100 years and manufactured the world's first electronic tdm switching system in 1970. Since then France's leading telecommunications company has installed its family of E 10 switching exchanges in 34 countries. In addition, the group has utilized its experience in digital switching to expand its line of switching systems to meet its customers' changing technical and commercial demands.

In an interview at CIT Alcatel's Paris headquarters, Christian Fayard, general manager of the public telecommunications division, discussed the E 10 switching system.

Scientific American: Why was a French company the first to develop an electronic tdm switching system?

Fayard: The R&D program we initiated with the PTT in 1965 was predicated on the belief that there would ultimately be a fully integrated electronic telecommunications network—combining digital pcm transmission with tdm switching techniques to increase efficiency, decrease costs and augment the communications capabilities of the telephone system. This was due to the increasing use of computers and microprocessors to perform data processing services—including call processing, circuit control and fault finding.

Our choice initially met with a great deal of skepticism. Most telecommunications specialists felt digital technology was applicable only in transit switching centers and were surprised when we began applying digital techniques to rural subscriber networks. Our first tdm switching system, the E 10, immediately illustrated the advantages of digital techniques over outmoded analog methods: better protection against noise, easier maintenance, improved fault detection and significant space savings.

Today, of course, tdm is the technology selected by most telephone administrations, and our international competitors are introducing switching equipment using this technique.

Scientitic American: What were the major achievements of the first E 10 system?

Fayard: The E 10 was the foundation for the rapid transition in France from an underdeveloped telephone network to one of the best in the world.Today France has an expanding modern network with 4.3 million E 10 lines in service.

Equally significant is the E 10's flexibility. The E 10 can be used to construct central office or transit exchanges. But it is an open-ended system and has been expanded to cope with the modernization of the French network in areas such as the electronic directory, direct satellite communication and cellular mobile radio telephone networks. For example, the E 10-Videotex is used as the access interface between the telephone network and the national videotex network. The E 10-Multiservice, used for the terrestrial switching centers of Télécom 1, incorporates the hardware and software modules required for processing the various terminal access protocols. It integrates the voice and data network by providing each subscriber with a high-speed data channel (up to 2 Mbit/s) that permits videoconference and other services. The E 10-Radiotelephone is used for a cellular mobile radio telephone network.



A CIT Alcatel minitel videotex terminal in the foreground of the E 10-Videotex switching system is an example of the adaptability of the E 10 to new telecommunications requirements. The unit provides the interface between the subscriber and the data bank, controls the connection, and handles format, transmission speed, and signaling conversions.

Scientific American: Hasn't the E 10 system become obsolete as the competition introduces new technology?

Fayard: No, because we have always kept the E 10 updated and upgraded by extending capacities and incorporating new components, such as vlsi microchips, 64 kbit dynamic RAM microprocessors and gate arrays. We are also introducing the international signaling code (CCITT No. 7) and common channel signaling.

It's been said that the E 10 is not a fully digital system, but it is in the American version, the E 10 Five, and will be in its 30-channel version with the introduction of CSN (the digital subscriber connection unit) in 1985. CSN is not only a single-channel codec but also the subscriber connection unit of tomorrow's isdn. However, the most important point is the system's total efficiency and reliability. Since its inception the E 10 has been function-dedicated, and today we remain the only distributed system operating on a large scale.

Scientific American: What are the emphases in R&D?

Fayard: The major R&D theme is the introduction of wide-band multiservice techniques. We are also studying the feasibility of optical switching units and are developing logic machines for software.

Software represents over half the development costs of a digital switching system (about 60 percent of the software is for operation and maintenance instructions and the remainder for function requirements), and we are aiming for the development of independent modular software that can be interlinked with processors using symbolic addressing. We do not believe there will be a new generation of switching systems but rather a continual evolution—a philosophy that is ideal for the E 10's modularity and distributed architecture.

Scientific American: How important is your international commercial presence, and what kind of contracts are necessary today?

Fayard: We have sold the E 10 system and transferred technology throughout the world. It is now manufactured in six countries, and our approach to future markets is typified by the large-scale project we are developing in India, which involves a complete transfer of technology and manufacturing know-how.

The E 10 is a very flexible system and can be used in most telephone networks. The E 10 Five, for example, has been adapted for the U.S. market as a central subscriber office.

based switched communications, which establishes a high-speed real-time connection between teletex-compatible terminals, word processors or electronic typewriters. Operating at a transmission speed of 2.4 kbit/s, it is 50 times faster than telex and produces documents in standard office letter format. Building on international standards, the French are preparing to launch a teletex service in 1984 to use the pstn and Transpac. It will capitalize on the French experience in teleprinter devices and the installed telex base in France, which includes over 100,000 subscribers.

The French are supporting international standardization of a mixed-mode teletex that incorporates digital facsimile capabilities. DGT officials argue that limited graphics are required so that texts can carry a signature and charts can be transmitted. Two firms, SINTRA Alcatel and SAGEM, are piloting the mixed-mode terminal applications, and a gateway to handle code, data rate and procedure conversions between teletex and telex networks should be operating by 1984.

"We are definitely pioneering standards in this area and are technically ahead in terminal identity procedures and the safeguarding of received information," says Philippe Leger, associate director of SAGEM, the world's second largest producer of telex terminals. "We will have a system that enables the transfer of graphics and terminal-to-terminal editing."

"France will develop a multimode teletex capable of transmitting written text and graphics on numerous data, telephone and telex networks," adds Roger Chaste, director of the telecommunications department at SINTRA Alcatel and a specialist in French telex display terminals and teleprinters.

The PTT is also studying electronic mail and electronic funds transfer within the postal system. One example of an existing service is Missive, marketed by France Câbles & Radio. Using a terminal connected to the telecommunications network, any subscriber can access a personal electronic mailbox through which he can prepare, exchange and file messages of all types to and from other subscribers.

VIDEOCOMMUNICATIONS: The government plans to order equipment for 1.4 million homes to be cabled in three years (100,000 in 1983, 300,000 in 1984, and one million in 1985) with a star system using multiservice fiber-optical local network transmitting images and sound. The cost is estimated at seven billion FF from 1983 to 1985, and manufacturers bidding for the work, scheduled to be contracted in late 1983, include CGCT, CIT Alcatel, SAT, Thomson-CSF and several smaller firms.

A second phase in 1986 will involve the expansion of the service to one million homes annually, and it is anticipated that one in every two French families could be connected to an optical fiber cable network in 1995. At that time the local networks will be interconnected through the high-speed national monomode transmission system.

"The goal is a multiservice wideband network with interactive characteristics employing fiber optics and a star structure," says Mexandeau. "The cable network, using the same architecture as the pstn, will increase capacity for existing services and promote videoconferencing, distributed data processing and distributed office automation."

In addition to the national plan, two tests are under way. In Biarritz at the

CGCT and its subsidiaries, LCT, la Signalisation and Pouyet, are today participating in every area of office communications. We draw on 100 years of experience to manufacture, install and maintain a wide range of telephone sets, public exchanges, digital PABX's, word processing, viewdata and data transmission equipment.

At our advanced research laboratories, LCT, we have developed a multiservice local area network, the Carthage Project. With its looped architecture, the network will be compatible with all types of voice, viewdata and data transmission equipment.

CGCT brings you the benefits of tomorrow's technology for today's communications.

CGCT

New concepts in office communications

CGCT 251 rue de Vaugirard,
75740 Paris Cedex 15
Tel: Area code (1) 545 2000
Telex: 200059 CGCT



CGCT's TLX 100 Interactive Word Processing Terminal

end of 1983, 1,500 subscribers will have access to 12 FM stereo channels, 15 TV channels and switched services that include videophone, videotex and data banks with image. SAT is the primary contractor and defined the system, and Thomson-CSF is an important subcontractor. In Lille, development of a local fiber-optical cable television system is under way for 50 subscribers, increasing to 3,000 in the 1984–85 period.

Mexandeau calls the French effort a "technically advanced, integrated, modular system that will rapidly become more competitive than conventional toll center distribution and telephone systems."

POSTAL SERVICES: The French have traditionally had an excellent postal service. During the next five years there will be increasing automation of the 17,200 post offices through the installation of microcomputers and minitel terminals. The PTT, which handles 14 billion pieces of mail a year, is studying money transfer, telecopying and new postage treatment centers. CIT Alcatel, Electronique Serge Dassault, SAT and other companies produce mail handling equipment, postage meters, automatic mail processing machines and electronic scales.

NETWORKS: France's nationwide 64 kbit/s digital network with end-to-end connections for transmission of speech, data and new services will soon create an itdn that will ultimately evolve towards the wide-band isdn—when the digital network will handle all services, including the transmission of moving pictures. The French are advanced in defining network architecture, network interface characteristics, user/machine access methods and net-

FRENCH COMMUNICATIONS:

TELECOM 1

The PTT uses INTELSAT, EUTELSAT and INMARSAT satellites for telecommunications, but the French also plan to launch their domestic satellite, Télécom 1, in April 1984. Matra is the prime contractor and is responsible for final integration of the satellite with the payload (including repeaters, transponders and antennas), being supplied by Thomson-CSF. Thomson-CSF and CIT Alcatel, through their joint venture Telspace, will supply earth stations, while Matra has a contract with the PTT for video transmission earth stations.

Télécom 1 is one of the key facets of the French multiservice communications system and will provide France with a spatial telecommunications network for the public and private sectors. While handling telephone traffic between France and its overseas territories and supporting video and defense communications, Télécom 1 will primarily guarantee efficient internal digital communications for French and European business traffic.

The business portion of the network, expected to have approximately 3,000 users, will utilize a maximum of 320 earth stations—with approximately 150 rural and 40 urban earth stations located in France. The quality, speed and cost-effectiveness of Télécom 1 has also led to the installation of additional earth stations in Switzerland, West Germany, Belgium and the United Kingdom in the framework of the European Telecommunications Satellite Organization (Eutelsat). Eutelsat will use part of the capacity of Télécom 1 for specialized intra-European services, while West Germany's Bundespost has contracted to use Télécom 1 to satisfy the digital private line needs of its national customers.

The applications payload of Télécom 1 permits the establishment of digital links of variable speed, from 2.4 kbit/s (i.e., text teleprocessing) to 2 Mbit/s (i.e., teleconferencing or high rate data transfers), between distant locations. This high-speed link, permitting numerous forms of transmission and processing of a wide variety and intensity of traffic, will assist the development of new types of communication requiring large bandwidths—such as video teleconferencing, electronic mail and fast-file data transmission between computers.

Télécom 1 offers the speedy installation of links, using earth stations with small-diameter antennae (3.5 meters), and an excellent adaptability to traffic growth. Most important, however, it provides a substantial degree of flexibility resulting from the utilization of the time division multiplex access (tdma) method of communication.

Tdma permits real-time demand allocation of digital channels and transmission capacity and allows users access to the satellite only when it is required, with the utilized capacity adapted to the service desired. The method is advantageous because the digitized signal can be regenerated on board the satellite, improving link quality, and the satellite transponder is used at maximum power, improving transmitter performance.

In the tdma access mode, each ground station transmits to the satellite on a common frequency occupying the totality of transponder capacity. Transmission ensures that blocks of digital data from different stations reach the satellite in sequence. This contrasts with the frequency division multiple access (fdma) technique, in which each station occupies part of the transponder capacity on a permanent basis.

A high-reliability central reference earth station located in Mulhouse in eastern France is used to control the network and on-request capacity assignment (the clock and frame synchronization) of traffic. This centralized intelligence will reduce the costs of tdma terminals in ordinary stations.



Matra

Télécom 1, the French multiservice domestic satellite, is being integrated at Matra's European test facilities. It is scheduled to be launched by the European Space Agency's Ariane launcher in April 1984.

Tdma experiments have been run with the Symphonie and OTS satellites, and the European ECS network will use this method exclusively. Specifications for the Télécom 1 system were prepared by the CNET with equipment supplied by a Thomson-CSF, CIT Alcatel and SAT consortium. The management of the Télécom 1 business system, designed by PTT affiliate France Câbles & Radio, will permit customers to have detailed records and information on communications expenditure.

France is also manufacturing TDF 1, a broadcasting satellite expected to be launched in the spring of 1985. The program involves a similar satellite for Germany; development and construction have been entrusted to Eurosatellite, in which Thomson-CSF and Aerospatiale are the French shareholders.

"There is an exciting future for digital business communcations," says Yves Fargette, president of France Câbles & Radio. "Télécom 1 will provide us with the know-how to expand our technology beyond France."

work control and management mechanisms.

In addition, CGCT has developed a fiber-optical multiservice local network for business offices, with an experimental version of the system installed at CNET research laboratories in Rennes. "Carthage" interconnects with the public telephone networks and, using the existing pbx, connects 600 telephone terminals, 400 terminals, and three computers, permitting the transmission of video, text, data and voice on one network.

French companies have also developed their own networks, such as Omeganet at SAGEM and Gixinet at Compagnie Internationale de Services en Informatique (CISI). Bull's Distributed System Architecture (DSA) is the central feature of its product strategy, and SESA is also studying the future approach to networking.

Some other interesting industrial developments:

—Audioconferencing has existed in France since 1976, and there are currently 100 public studios, known as *télécenters*, interconnected through the Caducée network.

—Telewriting permits the simultaneous transmission of written material using multiplexing techniques that allow input from a pressure-sensitive graphic table to be conveyed in the segment of the speech bandwidth between 1550 and 1950 Hz. France Câbles et Radio assembles the equipment, which is made by an affiliate of Télémechanique.

—Videoconferencing, achieved by transmission on a standardized 2048 Mbit/s channel on the telecommunications network, is under way in France, with 10 demonstration units throughout the country.

—SAT, SAGEM, SINTRA Alcatel, Thomson-CSF, TRT and other French and Belgian companies have developed the automatic integrated transmission network (RITA) for tactical military communications.

—A *Téléalarme* service, which is a home alarm system for elderly persons, has been inaugurated in Pau with ten alarm centers and 10,000 transmitters ordered from Thomson-CSF.

—Videotransmission permits the two-way projection of sound and images on large screens using microwave or satellite transmission. A network exists linking 13 large French cities, and mobile installations are also available.

RESEARCH AND DEVELOPMENT

The R&D effort in telecommunications technology is basically conducted under the auspices of the CNET, which particularly influences industrial development of telecommunications by preparing specifications, participating in consultations, monitoring market studies and standardizing new equipment.

CNET's total budget in 1983 was 685 million FF, with 3,827 persons working at six administratively decentralized research centers. General networks, optoelectronics and intercontinental links are studied in two Paris laboratories; local networks, switching, enhanced services, components and digital transmission are studied in two research centers in Lannion; new services and terminals are developed in Rennes at the CCETT (Common Center for Television and Telecommunications Research), and micro-electronics are studied at the Norbert Segard Microelectronics Laboratory in Grenoble. Research at CNET is largely responsible for tdm switching and pcm transmission, the tdma system for Télécom 1, the packet-switched data network and numerous *télématique* developments. Breakthroughs also include developments in solid-state physics, signal processing, radio communications and network design.

"Our work stops at the demonstration of feasibility, at which point industry takes over; our efforts today include a modernization of the current network to isdn, videocommunications and cable networking and research associated with the national electronics effort," says CNET director Jean-Pierre Poitevin. "We are digitizing the line to the subscriber, expanding the use of fiber optics with low-priced components, exploring new *télématique* orientations (such as voice recognition), improving submarine cables and conducting research on submicron micro-electronics."

Other research areas include future generation architecture and software for digital switching, liquid crystal display flat screens, second-generation videotex terminals, III-V semiconductor material, artificial intelligence and computer-aided design and manufacturing (cad/cam).

The CNET laboratory in Grenoble, for example, is developing a CMOS technology for telecommunications with the objective of producing chips using a 1-micrometer geometry by 1986. The laboratory is also working on strategic circuits for switching matrices, codec-filters and other applications related to videocommunication.

In Paris, network planning is being conducted at all levels (private, international, packet, telephone, videocommunication) to generate specifications for industry.

"We have a good lead in telephone and packet-switched networks, but not in data processing, which means that for local area networks (lan) France must often import a solution," says Jean-Marc Chaduc, director of the Paris A laboratory. "The most important aspect of any future network will be protocols."

"Our priorities are videocommunications, new electronic switching applications and the related software," adds Jean-Noel Mereur, director of the Lannion A laboratory. "We are digitizing the lan with combined data/voice communications and working on voice recognition and synthesis. In the latter area, we are advanced in algorithms but lack the necessary microelectronic components to create revolutionary developments."

CNET's network studies are leading to the digitization of a network integrating numerous transport options through the use of modular software.

"We look forward to basic specifications for the software to ensure adaptability for integrated systems that must support applications that are still not defined," says Daniel Hardy, whose R&D group in Lannion has an experimental multiservice switching system, known as Palme, under development as part of its general work on isdn.

The CNET is also conducting the Concerto software project aimed at demonstrating the possibility of a software workshop integrating the necessary tools for all software engineering activity—including design, development, documentation, qualification, maintenance and production.

"We are trying to construct models that provide the base and rationale for the next architecture of electronic switching systems," says Concerto project director Edouard André.

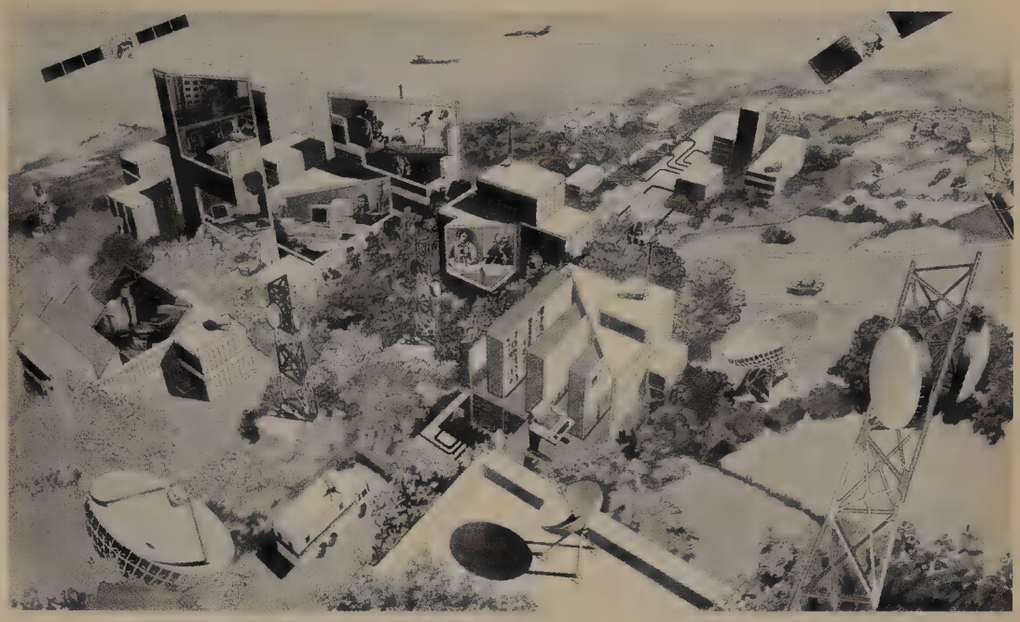
The CNET laboratory in Lannion began its R&D in fiber optics in 1970 and has since joined industry in installing numerous fiber-optical links throughout France. Today the CNET is using grooved cylindrical cable structures and establishing different multimode and monomode optoelectronic systems. The long-distance optical fiber link between Le Mans, La Fleche and Angers, for example, will employ both monomode and multimode fibers with a transmission rate of 140 Mbit/s, repeaters at 25 kilometers, and lasers emitting at the 1.3 μm wavelength. An advanced 42-kilometer monomode fiber link being tested in Brittany uses a 1.55 μm wavelength and has a rate of 4×140 Mbit/s.

"Like the rest of the world, we are leaning towards monomode networks," says Remy Bouillie, head of the optical networks division at Lannion B.

FRENCH COMMUNICATIONS:
THOMSON-CSF CONSOLIDATES ITS COMMUNICATIONS ACTIVITIES

The Thomson group, France's largest electronics company, with 82,000 employees and sales of 47 billion francs (45 percent in export markets) in 1982, has integrated the full range of its numerous communications activities. Thomson-CSF, the group's professional electronics affiliate, has formed a comprehensive communications branch to provide a complete approach to communications technology. Today Thomson-CSF is capable of handling all aspects of the acquisition, storage, processing, management and transmission of information in any form—voice, image or data.

The communications branch accounts for 25 percent of Thomson-CSF's activities, has 33,400 employees, exports 32 percent of its sales and allocates a substantial 14 percent of its revenue to R&D. It is involved in all aspects of telecommunications and electronics (components development, telecommunications, radio communications, satellite communications, office automation, networking and software), which form the foundation for the eventual integrated services digital networks. It develops thoroughly integrated products and guarantees the conception, manufacturing, installation and maintenance of turnkey networks and systems.



Thomson-CSF is France's complete communications technology company, with activities that include public switching, satellites, videocommuncations and business and military communications products and networks.

• Thomson-CSF is one of two French public switching equipment manufacturers and has 5 percent of the world market, with 20 million lines installed or on order in 20 countries (3.6 million digital lines to 16 countries). It has designed, developed, manufactured and installed exchanges based on electromechanical, space division electronic and time division multiplex (tdm) technology. The company now concentrates on its fully digital MT range of tdm switching equipment, which is one of the newest exchanges on the world market and can respond to the demands of a number of different public telephone requirements—urban, inter-urban, national, and international, with small, medium or large capacities.

Thomson-CSF has supplied turnkey telecommunications networks to clients including Chile, Egypt and Iraq and also encourages the transfer of technology. In the Soviet Union, for example, manufacturing facilities were constructed to produce one million MT lines per year. In addition, Thomson-CSF produces a full range of electronic pbxs, terminals and telephone sets.

• Thomson-CSF services the entire spectrum of satellite communications needs, with experience dating back 20 years to the first French and European scientific satellites. It has participated in virtually every European and several international satellite programs, is a member of the worldwide team working on Intelsat V advanced commercial communications satellites (providing transmitters, receivers and antennas) and will participate in Intelsat VI in collaboration with Hughes Aircraft Co. To date, Thomson-CSF has helped equip over 50 satellites and is the world's second leading supplier of earth stations.

Thomson-CSF is capable of manufacturing most spatial and terrestrial equipment for integrated state-of-the-art communications networks and systems. Its space-oriented activities range from components and traveling wave tubes to satellite payloads, onboard and terrestrial telecommunications equipment, and earth stations.

• Thomson-CSF is a major participant in the full range of transmission activities and is playing a key role in Biarritz, where France is introducing its first multiservice fiber-optical network. It is also installing the fiber-optical network in Lille and expects to participate in the national fiber-optical videocommunications program.

During the past 10 years Thomson-CSF has cut over a number of fiber-optical systems between public telephone centers, computers and railway stations for communication and surveillance requirements. It is capable of producing the full range of components required for a fiber-optical network—from cables to connectors.

Thomson-CSF's R&D activities embrace public switching, satellite communications and videocommunications, as well as a wide range of other disciplines in electronics and telecommunications. The group's R&D effort includes 18,000 persons, with 640 patents filed in 1982. In addition, Thomson has established specialized units to train foreign personnel and arrange technology transfers to other countries.

The company's current thrust is in the United States market, where it intends to expand its operations. GTE is selling Thomson-CSF communications terminals, 3M is marketing the Thomson-CSF desktop facsimile, and NBC and CBS are using Thomson-CSF's videotex know-how. With manufacturing subsidiaries or representatives in over 100 countries, Thomson CSF is playing a substantial role in telecommunications throughout the world.

THOMSON-CSF

173 BOULEVARD HAUSSMANN/75379 PARIS CEDEX 08/FRANCE TEL. (1)561.96.00/TELEX: ELIHU PARIS 650143. TCSF 204780F


PTT


TRT


Thomson-CSF

The minitel terminal is being used in numerous businesses and homes in France and has been adapted to other languages for export sales. Thomson-CSF has sold its communication terminal to GTE in the United States.

"We are also testing the feasibility of plastic fiber, optical switching and a local optical network."

The prototype of the CNET's local optical network—including transmission, signal processing and switching—employs a single multimode fiber to offer bidirectional service.

THE FOREIGN PRESENCE

Not all foreign companies can be expected to be pleased with French policies. ITT's French subsidiary was nationalized, and National Semiconductor sold its stake in Eurotechnique to Thomson-CSF "for a symbolic one franc." But numerous foreign firms operate in France, including IBM, Philips, ITT, Motorola, Hewlett Packard, SGS Ates, Texas Instruments, Wang and Tektronix.

IBM is the major participant, and its telecommunications products include pbxs, modems, multiplexers, telex software, and special adapters for packet-switched networks. Its international telecommunications research center is located at La Gaude, near Nice.

"Our major concern about France is the threat of protectionism and the direction of the economy," says Hervé Caron, assistant general manager of IBM France. "Although we certainly don't depend on the French government for our general strategy, we are an integral part of French electronics efforts and are reasonably confident about our future."

American firms feel they should take advantage of the French position in telecommunications.

"We gained tremendous insight into packet-switched data networks, videotex, fiber optics and the trend toward isdn because of our presence in Grenoble," says Wim Roelandts, R&D manager for Hewlett Packard's information products group. "There has recently been a much more open attitude by the government to establish joint R&D projects with us."

"We are pleased with our agreement with Matra, which continues to have the government's blessing," adds Jean-Claude Rivet, Intel's assistant general manager for Europe. "We do not have enough design resources in the U.S., and that joint venture provides us with European expertise."

TRT, in which the Dutch Philips group has a 45.6 percent stake, uses that link to market French innovations on a worldwide basis.

"Obviously we will use the Philips distribution network to sell the electronic directory in Europe," says TRT commercial director Marc Houery.

Other foreign companies are rapidly expanding in France. The world's first commercial Wangnet has been installed in the Citibank headquarters in La Défense, and Wang may construct manufacturing facilities in Normandy. Apple Computer has reopened negotiations with the French government to locate its European software center in France and may also distribute French-manufactured hardware in the U.S.

COOPERATION

The French are playing a role in mounting a European challenge to American and Japanese predominance in electronics and telecommunications, and the PTT is attempting to establish pan-European technical and commercial ties.

"We recognize the importance of IBM in our industry," says Mexandeau. "Yet at the same time we must fortify our European connections if we are to adequately compete with Japan and the U.S."

Obviously, a European solution could brighten the future of French electronics and telecommunications. The 10 European Economic Community (EEC) countries represent 30 percent of the world market for information technology, but manufacture only 15 percent of the products. Cooperation would provide a larger home market, standardized products and more export clout. One French-supported venture in this direction is the European Strategic Program of Research and Development in Information Technology (ESPRIT), which has a goal of tripling the part of the EEC in world markets by 1990.

There is also a possible pan-European collaboration to develop future switching equipment, and Bull is discussing joint research programs in the data processing field with West German and British companies.

"We have defended our local market against foreign penetration without protectionist measures and have proven our technical and commercial strengths abroad," says CIT Alcatel's Chavance. "We still hope that we can reach important agreements with European partners."

EXPORTING FRENCH PRODUCTS

France is the fifth largest telecommunications exporter behind West Germany, Japan, the United States and Sweden. Orders for telecommunications equipment increased from 5.4 billion FF in 1981 to 6.3 billion FF in 1982, with exports representing 24 percent of total sales. The main export sectors were switching (36 percent) and transmission equipment (27 percent), and the principal markets were the Middle East (29 percent), Africa (27 percent), Europe (27 percent) and Latin America (7 percent).

The government acts as a vanguard for the promotion of most French products abroad. France Câbles & Radio, for example, has subsidiaries in numerous countries to construct submarine cable systems and manage international telecommunications activities. More important, when French President Mitterrand and officials travel, telecommunications is often on the agenda during high-level discussions, and French banks are extremely cooperative in financing PTT-backed export projects.

"We have a variety of products and services that are capable of penetrating international markets, and it is our job to promote good French equip-

FRENCH COMMUNICATIONS:
SPEARHEADING ELECTRONIC FUNDS TRANSFER

The French PTT, banks and industrial companies have united to launch an on-line point-of-sale system (OLPOSS) that has the potential to penetrate export markets, thanks to its economy, flexibility, security and simplicity. The project combines data processing, telecommunications and financial networks to create state-of-the-art electronic funds transfer (EFT).

The installed system allows a complete financial transaction, including authorization and recording of data, to take place in only a few seconds. It makes it unnecessary to transmit *any* additional information between the customer, the retailer and the bank. The one-step process involves magnetic stripe credit cards supplied by French banks and a multiservice terminal connected to a central computer.

An initial 30-month test in Saint-Etienne, known as the Point Rubis trial and estimated to cost 30 million FF (including design, research, investment and operating costs), is being financed by a consortium of French banks, the PTT and retailers. Saint-Etienne was selected because it is a large regional center (population 220,000) with an extensive business infrastructure.

Jacques Mayoux, chairman of Société Générale, the world's ninth largest bank, which is coordinating the project, contends that "the experiment has been successful because of the cooperation between oft-competing financial institutions, retailers and industrial partners."

The on-line procedure includes entry and validation of transaction data, transmission, data checking and data recording. Once a user's credit card is inserted into the terminal card reader and purchases are totaled, the user employs a hand-held key pad to enter the required confidential authorization information—a personal identification code (PIN). The terminal then establishes communications with the Interbank Network Management Center to verify the PIN, accept or deny the transaction and record the data needed to credit the retailer's account and debit the customer's account.

Data is sent from retailers' terminals to either a local concentrator or a public switch concentrator installed in a telephone exchange. The concentrator acts as a packet assembler/disassembler for transmission to the central computer via Transpac, the French public packet-switched data network.



The public point-of-sale test in Saint-Etienne uses a multiservice terminal and credit cards with magnetic stripes to interact in real time with a central computer over the French packet-switched data network. The user is providing the system with her personal identification code.

The clearing of payment between banks, crediting of retailers' accounts, debiting of customers' accounts and transmission of daily transaction statements to retailers are handled in batch mode from the central processing unit.

The prime contractor for the system is Société Générale's software engineering and information services company, SG2 (Société Générale de Service et de Gestion). SG2 has provided software for a number of financial networks and systems and is responsible for all the software, codes and algorithms in the central computer.

Terminals and concentrators were designed and manufactured by Electronique Serge Dassault, which is a specialist in producing financial and control terminals. The 5-kilogram terminal for Saint-Etienne includes a half-duplex modem, a magnetic track reader, a keyboard and display unit for the retailer, a printer (for the customer's charge slip and the retailer's records) and a keyboard and auxiliary display unit for the customer.

"There was a successful synergy between the talents of a software house and a hardware house to produce a fusion of systems," says Serge Dassault.

Electronique Serge Dassault and SG2 are commercializing the EFT system through the joint venture company OLPOSS. They believe there is a significant export market, since the system is easily transportable and can be delivered as a turnkey network in less than two months. Software can be modified to suit local requirements, and the system can be adapted for numerous other uses—such as purchasing transport and parking tickets or paying highway tolls.

The benefits of OLPOSS are numerous. For retailers the system provides speed and security while eliminating paper and the risk of unpaid bills. For banks, the cost per transaction is about five times less than with checks and half that of a regular credit card transaction. While OLPOSS is not expected to eliminate traditional payment methods, it is expected to stabilize the use of checks in France. An adapted system will be used for home banking and telepurchasing applications in conjunction with domestic videotex terminals.

"The key to the system is the security of transmission, the rapidity of interrogation, and the efficient routing of information," says Philippe Tournaud, associate director of SG2. "In addition, the system can easily be expanded to take new types of cards—such as smart cards or cards employing holographics."

Numerous foreign banks and financial companies have expressed interest in the system, which is also available in an off-line mode. More important, French banks, public authorities and industry have developed a common strategy that will spearhead the implementation of EFT and establish a French lead in this high-growth technology.

GROUP SG2

12, AV. VION-WHITCOMB
75016 PARIS
TÉL. : (1) 524.52.22

55 QUAI CARNOT.
92215 ST-CLOUD CEDEX. FRANCE.
TEL: (1) 602.50.00

ment," says Jacques Dondoux.

To facilitate export of the French videotex products, the French are adapting to local market requirements. The database consultation dialogue of the electronic directory, for example, has been translated into English, German, Spanish and Arabic, and videotex systems have been sold in Brazil, Kuwait, Italy, Australia, New Zealand, the United States and Greece.

Naturally, the large, expanding U.S. telecommunications market is important terrain for the French.

"No French company has an excellent base in the U.S., which is where the most important marketing battles will be fought in many areas of telecommunications," says SAT's Le Menestrel. "It's the number one priority for every French company."

SAT, like most French companies, has established a variety of agreements and ventures with U.S. firms. For example, it took a stake in IPC Communications to market its pbx equipment and has a joint venture with General Optronics for fiber-optical products.

CIT Alcatel has numerous joint programs and cooperative ventures. Lynch Communications System, in which CIT Alcatel has a 25 percent share, has given it a U.S. base for marketing the E 10 switching exchanges and the electronic directory. CIT Alcatel recently won a major contract in India and will build two local plants to manufacture E 10 digital switching exchanges, each with an initial capacity of 500,000 lines.

"The Indian agreement is an ideal transfer of technology," says Mexandeau. "It is the type of exemplary agreement we would like to pursue in selling French products."

Thomson-CSF is active throughout the world, with major contracts in Russia, Iraq, Lebanon and the Gulf states. Its U.S. activities are conducted by Thomson-CSF Communications Inc. "We will be very active in the U.S.

FRENCH COMMUNICATIONS:

SUPPLYING ENGINEERING EXPERTISE WORLDWIDE

Exporting French telecommunications expertise and engineering techniques is a government priority that has been underlined as a result of the wide range of advanced products and programs recently introduced in France. French telecommunications experts offer their know-how in telephone network design, maintenance and operation methods to public administrations and companies throughout the world.

The DGT also assists foreign administrations in developing and expanding their training facilities. One third of all DGT experts sent abroad do so within the framework of technical cooperation. Foreign technicians visiting France are provided with the training—in the DGT's engineering schools, in specialized training centers or in specific seminars—required to become acquainted with French techniques and equipment. And the CNET organizes numerous scientific and technical exchanges between foreign research institutes in all fields of advanced or applied research. The result is a two-way street in the transfer of technology between France and developed and developing countries.

Another important participant in this international role is SOFRECOM (Société Francaise d'Etudes et de Réalisations d'Equipments de Télécommunications). Established in 1966, it provides consulting and engineering services throughout the world and is independent of any industrial or commercial group. The PTT has a 32.8 percent stake in SOFRECOM (the remaining shareholders are seven French banks). SOFRECOM's revenue has increased from 23.5 million FF in 1979 to 55 million in 1982—with over 80 million FF forecast for 1983.

The organization has over 300 contracts in more than 40 countries following international bidding competition between major consulting firms. It is approved by the International Telecommunications Union and major international and regional development banks.

SOFRECOM provides four types of services: planning (from demand forecast to network planning); engineering at all stages of design; technical assistance for operation and maintenance; and management consulting. SOFRECOM's staff of 90 (40 of whom are engineers) are specialists in all the required switching, transmission, power and data processing areas associated with telecommunications. Most of the experts have been active in the PTT, where they gained firsthand experience in the implementation and operation of telecommunications equipment.

An example of SOFRECOM's role in international technical cooperation is the general assistance it is providing for the modernization of India's telephone network. SOFRECOM assists the Indian telecommunications administration in mastering the introduction of digital exchanges in an analog network with the help of management methods and tools developed by the French administration. The contract involves training in France and India and the supply of engineering, testing, operation, maintenance and managing methods.

SOFRECOM also provides consultancy services for a complete regional transmission network, including traffic studies, cable and microwave route surveys, drafts of specifications, tender evaluations, assistance during contracting and acceptance testing.

In addition to SOFRECOM's activity, France Câbles & Radio acts as a consultant in international telephony for public telecommunications networks, satellite communications, radio communications, teleconferencing, submarine systems, computer systems and data transmission networks, private networks and office automation. It also forms operating companies for international telecommunications and submarine systems in association with foreign governments. Its affiliate Intelmatique is the French promotional and marketing organization for the smart card and videotex products, and another subsidiary, France Telecom, promotes French telecommunications in New York, Singapore and Caracas.





French companies have exported their communications equipment throughout the world. Thomson-CSF installed a complete turnkey microwave network in Zaire, and SOFRECOM was responsible for the design of a microwave network in the Middle East.

during late 1983 and throughout 1984 in capital risk, financial participation, joint ventures and other forms of collaboration," says Jacques Darmon.

Matra is a major exporter and is often associated with other important international companies: TRW and British Aerospace for space projects, Harris Semiconductors and Intel for integrated circuits and Datapoint for data processing. Matra and Tymshare, the California-based computer services company, have formed Tymshare-Matra Corporation (TMC) to distribute 560,000 Scanset terminals in North America. Tandy's TRS 80 Model III minicomputers are being manufactured in Colmar as part of the agreement with Matra.

SESA is creating Auspak, the X.25 public packet network operated by Telecom Australia, with the same DPS 25 packet switch that it is using for public networks in Brazil, Luxembourg, New Zealand, the U.S. and Europe.

"We obtain 50 percent of the contracts awarded for packet-switching data transmission networks throughout the world," says SESA president Jacques Arnould.

TRT has recently received contracts for microwave networks in the Congo, Saudi Arabia, Nigeria, the Ivory Coast, Zaire, Jordan, Lebanon and Tunisia. It also recently won a contract to supply one hundred IRT 1500 stations throughout New Zealand for the rural telephone system.

Exports account for 28 percent of Jeumont-Schneider's telecommunications revenue, and the JISCOS switching equipment and range of pbxs have been exported to 30 countries. TIE Communications Inc. distributes Jeumont-Schneider products in the United States, and the company is developing new products for the U.S. market.

The French also hope to catapult their software expertise into an international presence. CISI took over the Wharton Econometric Forecasting Associates as part of its long-term strategy to develop an international data processing network. Steria has been asked by an American bank to develop new services for its videotex system, including integration of the smart card, which has considerable export potential.

Videographic Systems of America (VSA) has been formed by Thomson-CSF (51 percent), Cap Gemini Sogeti (13 percent), Steria, and other French companies (government organizations hold 16 percent of the capital) to penetrate the U.S. market. CIT Alcatel has formed a marketing agreement with SESA and Cap Gemini Sogeti, and Matra and Télésystems have created a

A WORLD LEADER IN DATA PROCESSING ENGINEERING AND TELECOMMUNICATIONS

SESA is the foremost designer and manufacturer of:

* ***data transmission networks*** *– a major contractor for public and private networks in France (Transpac), Brazil, Australia*

* ***turnkey télématique systems*** *including the electronic directory and telephone information systems*

* ***real time data processing systems for:***
  - *ticket automation for subway systems in Paris, Baltimore, Caracas and other major cities*
  - *dispatching electricity*
  - *satellite monitoring*
  - *military information systems*

SESA: An international company with over 1200 employees which has doubled its sales during the last two years.

SOCIETE D'ETUDES DES SYSTEMES D'AUTOMATION
SESA, 30 Quai de Dion Bouton
92806 PUTEAUX – FRANCE
Tel: 776.41.00 Tlx: 610022F

SESANET Inc., 11410 Isaac Newton Square,
RESTON, Virginia 22090, U.S.A.
Tel: (703) 435.88.90 Tlx: 7108331176 TSS RSTN

SESA

LEARN MORE ABOUT FRENCH COMMUNICATIONS

Advertisers in this special *Scientific American* advertising supplement will be pleased to provide additional information about their products and services.

☐ Bull
☐ CGCT
☐ CIT Alcatel
☐ DGT/DAII
☐ Intelmatique
☐ Jeumont-Schneider
☐ SESA
☐ SG2 Electronique Serge Dassault
☐ Thomson-CSF

Interested readers should tick the appropriate box and send the completed coupon to: Manager—Industrial Sections, *Scientific American*, 415 Madison Avenue, New York NY 10017, USA.

Your request will be forwarded to the appropriate company. They will send material directly to you. Please allow time for processing your request.

Name .....
Title .....
Company .....
Address .....
City .....
Country .....



A researcher at the CNET laboratory in Lannion in western France is using Mastermind to test voice recognition techniques (left) and can also instruct the system to draw lines (right) connecting various French cities. CNET.

joint venture for the exportation of interactive videotex systems.

"The French products come from the same environment; thus uniting to make the technical and commercial adaptations for the American market is extremely sensible," says Bernard Joseph, president of VSA.

"VSA has produced a general awareness of the French presence in this area and shows that the French, like the Japanese, can adapt their products to American standards," adds Gregory Harper, vice president of VSA. "The joint venture also reveals that the French realize a group approach is sometimes necessary to penetrate the U.S. market."

However, "We don't need a French 'get together' just to sell the electronic directory in the U.S.," counters CIT Alcatel's Chavance.

MARKETING FRENCH EQUIPMENT

Although the French have technical leads in some sectors, there are a number of hurdles to overcome before they can be considered masters of international telecommunications. The serious economic situation in France could retard ambitious national programs and the resulting introduction of low-priced products on international markets. New projects, such as the fiber-optical cabling of France, are extremely ambitious in an era when financing is limited and the technology is untried.

There are also some ingrained difficulties. In measuring equipment and instrumentation, for example, there is what one government official calls "the traditional *need* to buy American." And the French have been attempting to create a solid computer industry since the 1960's without much success.

"We have not performed well in the past because there hasn't been a coherent long-term plan," contends Jacques Stern, president of Bull. "We now have that plan, we will implement it, and it will succeed."

Such assertions notwithstanding, the French must become more realistic. Their ambitious electronics plan is too expansive and too expensive in its present format.

Most French companies also lack effective worldwide commercial strategies, and the government sometimes seems more intent on boosting domestic employment than on increasing exports. The emphasis should be reversed, and the French should move quickly to establish a greater international presence.

Although the French have alienated some potential multinational partners as a result of their nationalization program, they are wise to support more cooperative ventures—in Europe, the U.S., Japan and the Third World. Future bilateral and multinational agreements could provide the commercial backbone that would enable French companies to profit from their established technological lead. Collaborative ventures are definitely required if the French are to penetrate newly deregulated European markets—something they have been slow to do in the past—and the transformed American marketplace.

"Telecommunications in France is a relatively new industry, and we still don't have the commercial astuteness of some of our competitors," says Mexandeau. "But we are initiating new products far ahead of the competition and must capitalize on our lead by selling throughout the world. Our results are good in general; however, we are virtually still on square one in some markets."

To offset their lack of commercial acumen, the French nationalized banks do deserve credit for financing large export projects supported by the PTT. This financial element is extremely important in international telecommunications, particularly when the client is a government administration, and the French have the grudging admiration of their competitors in formulating attractive financing packages.

Nevertheless, they cannot rely on a captive domestic market and the supportive banking industry. They understand that innovative marketing strategies, including more commercial agreements, are required to sell their telecommunications equipment on an international scale. One indication that they are improving their commercial role is that their national exhibition at the quadriannual "Telecom 83" in Geneva in late October is seven times larger than it was in 1979.

"We must start moving faster than the Japanese," concludes CIT Alcatel's Chavance.

Front Page Photograph. The French government and industry have initiated a number of new communications products, systems and services, including microchips, low-cost terminals, fiber-optical networks and smart cards. Credit: GILBERT LEGAY

FRENCH COMMUNICATIONS:
BULL: THE INTERNATIONAL INFORMATION SYSTEMS GROUP

France is concentrating its expertise and resources in the area of information processing systems around Bull, a group which has been internationally active in the data processing industry for fifty years and today successfully markets its products and systems in over 75 countries. Bull, with 25,000 employees and a 1982 revenue of 10 billion FF, is a key element in the French *télématique* industry. The group has long acknowledged and supported the evolving integration of telecommunications and data processing techniques. It is playing a major part in the country's long-term thrust in information technology.

Bull, majority-owned by the French government, has announced its technical and commercial strategy for the coming years designed to strengthen its position notably in distributed data processing, business communications and office automation. The group now includes Cii Honeywell Bull, minicomputer manufacturer SEMS, distributed data processing company Transac and microcomputer producer R2E. New divisions have been formed with full responsibility for planning, development and manufacturing: medium and large mainframe systems (Bull Systèmes), minicomputers (Bull Sems), peripheral equipment (Bull Périphériques) and office automation and microcomputers (Bull Transac). All products and systems are marketed internationally through a single sales organization.

Increased capital investment will enable the group to realize its goals. Bull will also benefit from its mainframe and networking experience and develop, at the international level, new applications and activities in distributed data processing.

The French government, assuming its responsibility as main shareholder, invested 1.5 billion FF in Bull during 1983—thus providing the resources to balance the insufficient capital investment of the last five years and implement the group's new strategy. The group projects an annual growth rate of about 16% concentrating on four major market sectors:

1. Bull is consolidating and developing its installed base in medium and large data processing systems.
2. Bull is penetrating the distributed data processing market with a range of minicomputers, microcomputers, office automation equipment and terminals based on a global systems and networks approach. The interweaving of the traditional fields of office equipment, telecommunications and data processing and the overlapping of microcomputers, minicomputers and mainframe systems are key elements in the group's approach.
3. To support these two lines of action, Bull is cooperating with software houses and users to expand the supply of services, such as software packages (specific applications software aimed at specific market segments), technical assistance, education and training.
4. Bull is also expanding its computer peripheral activities on the OEM market in areas including compact disk drives and non-impact magnetographic high-speed printers.

Bull will also capitalize on its established position in rapidly expanding areas such as networking. Distributed Systems Architecture (DSA)—the Bull network which offers a unified, coherent approach to data communications within a distributed system—implements all internationally agreed protocols and allows Bull products to connect directly to most existing public and private networks. DSA is an "open" architecture which enables the group's products to cooperate with products from other suppliers operating in other types of network such as SNA. DSA enables Bull to meet expanding demand in this market and, to ensure a solid marketing position in the future, development projects are underway to extend this architecture to local area networks and electronic message and mail applications.

While Bull will naturally develop its current product lines and use them as a springboard for expansion, it is also active in innovative areas.

The Bull CP8 smart card, for example, is a plastic card which contains a self-programmable microprocessor and memory chip. It can be used as a means of payment, an access key, or a portable file of personal information. Bull is making a considerable research effort in this field and possesses a technological lead due to pioneering R&D and experimentation with new applications. Bull's CP8 smart cards are being used in numerous pilot projects in Europe and throughout the world for point-of-sale payment, remote purchasing/payment, home banking, public payphone operations and personal record keeping.

Other examples of innovation can be found in such widely differing hardware and software domains as non-impact printer technology and program language development. The MP non-impact high-speed printer uses a new magnetographic technology developed in-house to combine reliability and high quality with low cost. The ADA programming language, developed by the group for the US Department of Defense as a replacement for the many incompatible languages now in use, is gaining wide industry acceptance.

Bull's relatively ambitious strategy relies on the continued strengthening of worldwide technical and commercial cooperation. Bull has built up one of the largest international marketing networks in the computer industry—an asset that will facilitate future growth. In France, Bull has increased its systematic cooperation with French universities, public research bodies and industrial companies. For many years, Bull has cooperated closely with Honeywell, which holds a minority stake in the French group. This association, which has resulted in common product lines, was extended last year with a new set of long-term technical and marketing agreements. As part of its policy to promote cooperation, Bull also acquired a minority stake in the capital stock of Trilogy when that computer company was founded by Dr. Gene Amdahl in 1981. The group also has commercial and technical agreements with other American companies such as MPI and Convergent Technologies. It also collaborates with European companies and research organizations in a variety of areas.

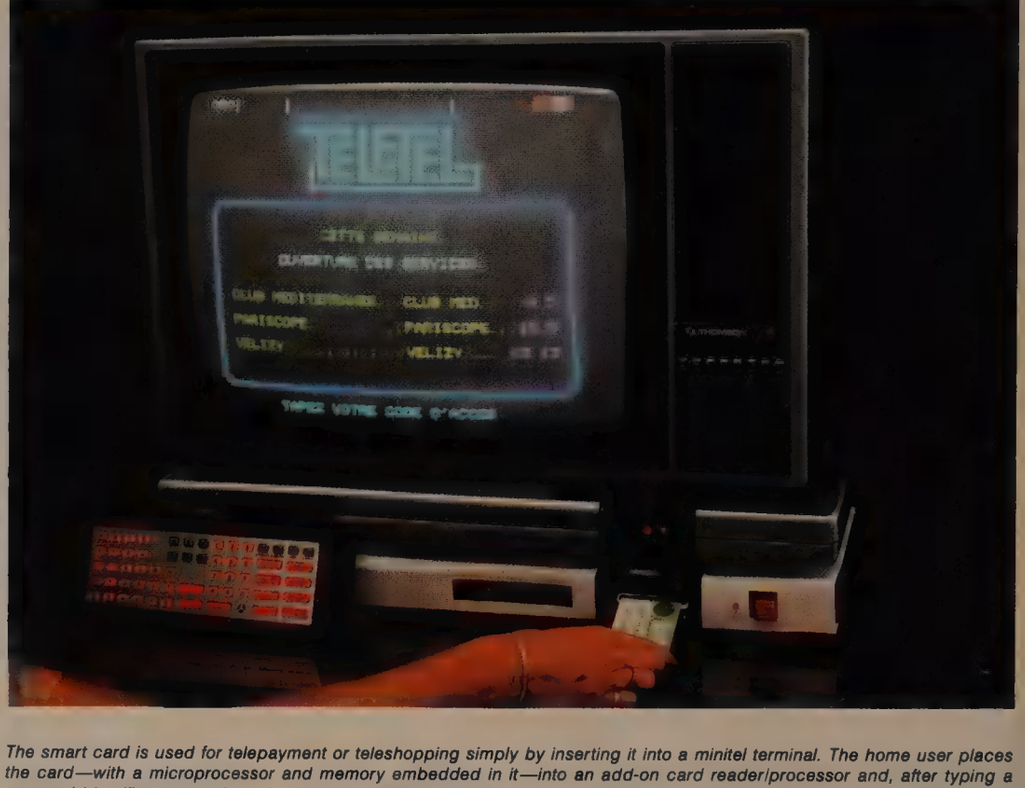
R&D is an important sector with a 1983 budget of 1 billion French francs. While the group is making headway in augmenting the capacities and capabilities of its existing products, it is also conducting basic researh into magnetic storage and non-impact printer technologies, fiber optics to create high-performance communication channels between computers, artificial intelligence, software engineering, logic machines and the development of new integrated circuits using custom vlsi technology.

These combined activities make Bull a strong international information systems group which guarantees advanced technology and continuity in the development of an integrated range of computer and communications products for today and tomorrow.



HOW CAN YOU FIND OUT MORE ABOUT THE SMART CARD AND VIDEO TEX PRODUCTS?

Just ask Intelmatique. We're the international promotional, marketing and consultancy company for *télématique* products. In fact, we're the ones responsible for coining the phrase "smart card" and launching a number of Teletel videotex projects internationally.



The smart card is used for telepayment or teleshopping simply by inserting it into a minitel terminal. The home user places the card—with a microprocessor and memory embedded in it—into an add-on card reader/processor and, after typing a personal identificaton number, the memory written into the card by the issuing authority identifies the user, indicates the money value or banking power of the card and automatically dates and records information within the integrated circuit. The card can be used for a variety of other banking and business applications.

As the official representative of the French Telecommunications Administration, we're in a unique position. We not only monitor *télématique* activities throughout the world but are in constant contact with French *télématique* programs, research projects and industrial developments. We can provide you with details about proven applications and technical specifications for the smart card and other products. Why? Because we've overseen the introduction of the smart card for numerous user trials from Fort Lee to Oslo and studied the consumer and banking reaction to point-of-sale experiments.

We also can supply you with the know-how gained from six years of international experience in the videotex field. We've held seminars on the smart card and videotex from Chicago (where we've opened a public videotex demonstration and support center in conjunction with Honeywell) to Australia. We've helped install videotex trial systems in Kuwait, Sao Paulo and Minneapolis. We've put videotex terminals on executives' desks, in middle size businesses, in bank branches and in living rooms. Perhaps you haven't heard about the French range of low cost terminals (minitels) which can handle information while interfacing with your own system. These inexpensive terminals could be the answer to your in-house communication needs and can be used in conjunction with the smart card.

If you are ready to evaluate the smart card or other videotex products let us open the doors. Intelmatique—the international promotional and marketing arm of the French PTT—has stimulated interest in many countries and our widespread experience can help you. We can show you the technology available today and planned for tomorrow. We are also able to offer professional consulting in most parts of the world if your need is for expert assistance in developing a business strategy.

To join the smart set and make the link with télématique simply mail us this card. We'll send you additional information.

Yes, I'd like to receive additional documentation on

The smart card ☐ or videotex ☐.

Name .....

Position .....

Company .....

Address .....

City .....

Country.....

Intelmatique
FRANCE

THE ORGANIZATION FOR THE INTERNATIONAL PROMOTION OF TELEMATIQUE
98, rue de Sèvres 75007 PARIS. Tél. : (33) 1 306.16.36, Télex 203185 TELEMAT. F.

or call our representatives in U.S.A. : (203) 834.11.43; Australia : (61) 2. 27 21 81; New Zealand : (04) 844.870; West Germany : (49) 611.683271; Greece : (30) 1.9228.

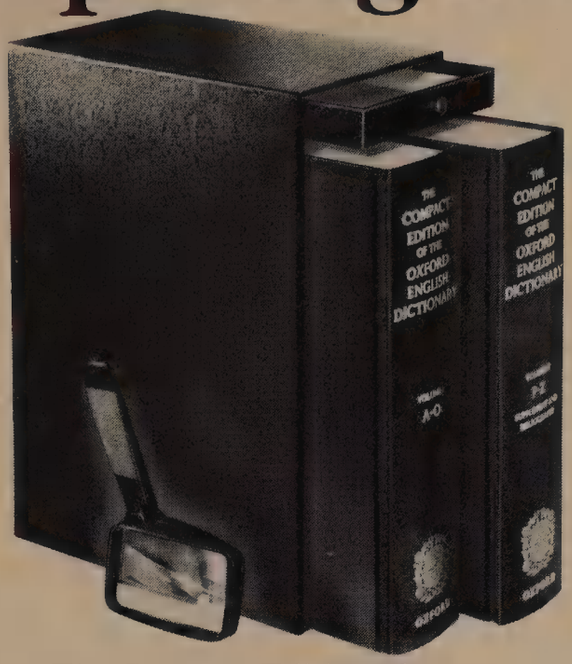
Sets of permanent value at surprising savings.



The Encyclopedia of Philosophy

for **$24.95** (Pub price $175)

The most comprehensive encyclopedia of philosophy ever published, this set traces ancient, medieval, modern, Eastern and Western thought. An essential and rewarding reference source for home libraries.



The Compact Edition of The Oxford English Dictionary

for **$24.95** (Pub price $150)

The "most complete, most scholarly dictionary of the English language"—*The Christian Science Monitor*. Through photoreduction, the original 13-volume set has been reproduced in this two-volume *Compact Edition*. Magnifying glass included.



The Nine Beethoven Symphonies & The Five Piano Concertos/Ashkenazy

for **$19.95** (List prices total $115.74)

Two of the most celebrated recordings of Beethoven's works. Sir Georg Solti conducts the Chicago Symphony in the nine-record set of Beethoven's complete symphonies. "A glorious musical combustion"—*Time*. In collaboration with soloist Vladimir Ashkenazy, Solti and the Chicago Symphony perform Beethoven's five concertos. "A major addition to the Beethoven discography"—*High Fidelity*. Available in 13 records or 8 cassettes.



The Decline and Fall of the Roman Empire by Edward Gibbon. Edited by J. B. Bury

for **$24.95** (Pub price $300)

The definitive Bury edition of the most acclaimed history of all. Gibbon evokes a world of grandeur and decadence, masterfully tracing its collapse under emperors noble and ignoble. Seven-volume set, newly available with maps and illustrations on long-lasting, acid-free paper quarter-bound in leather.



The Story of Civilization by Will and Ariel Durant

for **$29.95** (Pub prices total $335.45)

For almost half a century Will and Ariel Durant traced the continuity of world history—the religions and philosophies, the political and economic tides, the arts and sciences, the customs and conquests—to show the foundations of society today. A Book-of-the-Month Club exclusive for almost 50 years, the Durants' illustrated masterwork is history come alive.

Choose one of these five sets and save up to $305.50

You simply agree to buy 4 books within the next two years.

Facts about Membership. As a member you will receive the *Book-of-the-Month Club News*® 15 times a year (about every 3½ weeks). Every issue reviews a Selection and 150 other books that we call Alternates, which are carefully chosen by our editors. If you want the Selection, do nothing. It will be shipped to you automatically. If you want one or more Alternates—or no book at all—indicate your decision on the Reply Form and return it by the specified date. *Return Privilege*: If the *News* is delayed and you receive the Selection without having had 10 days to notify us, you may return it for credit at our expense. *Cancellations*: Membership may be discontinued, either by you or by the Club, at any time after you have bought four additional books. Join today. With savings and choices like these, no wonder Book-of-the-Month Club is America's Bookstore.

BOOK-OF-THE-MONTH CLUB

America's Bookstore® since 1926.

Book-of-the-Month Club, Inc., Camp Hill, Pennsylvania 17012 **A170-10**

Check one box only.

- ☐ 912. Compact OED **$24.95**
- ☐ 913. The Story of Civ. **$29.95**
- ☐ 917. Ency. of Philosophy **$24.95**
- ☐ 932. Beethoven (13 Records) **$19.95**
- ☐ 933. Beethoven (8 Cassettes) **$19.95**
- ☐ 951. Decline and Fall of the Roman Empire **$24.95**

Please enroll me as a member of Book-of-the-Month Club and send me the set I have checked at left, billing me for the appropriate amount, plus shipping and handling charges. I agree to buy four books during the next two years. A shipping and handling charge is added to each shipment.

Name_____ 3-64

(Please print plainly)

Address_____ Apt._____

City_____

State_____ Zip_____

Prices shown are U.S. prices. Outside the U.S., prices are generally higher.

The Processing of RNA

DNA is transcribed into messenger RNA, which is then translated into protein. In cells that have a nucleus, including human cells, a lot happens to the RNA between transcription and translation

by James E. Darnell, Jr.

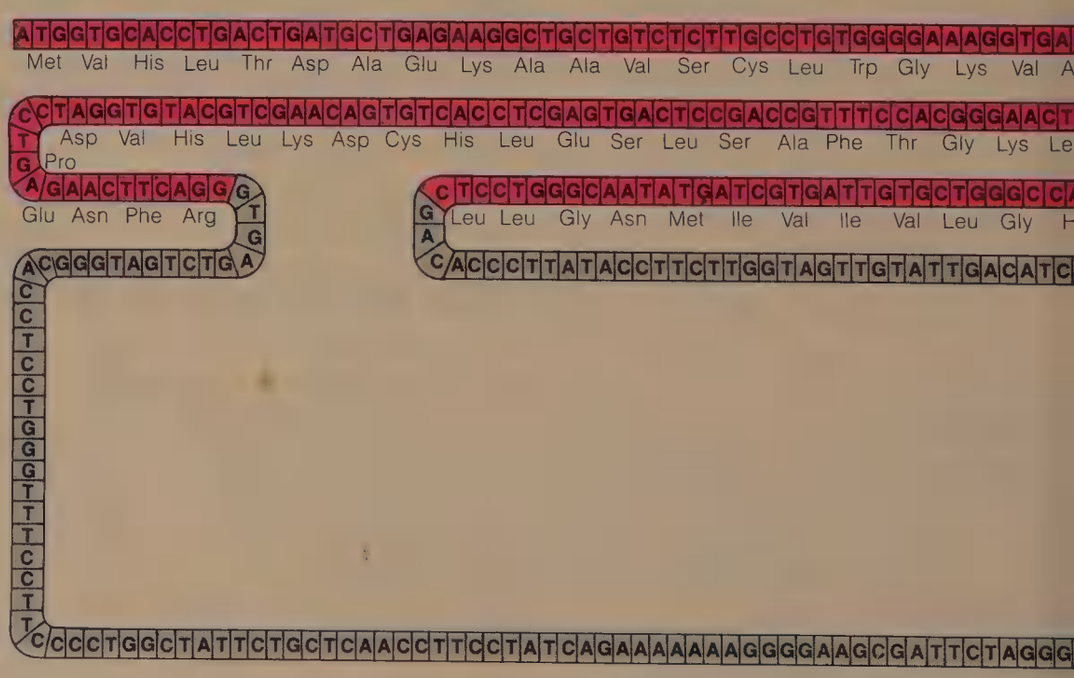
One of the central problems of molecular biology is to discover how two cells with the same set of genes can each make a different array of proteins. It is probable that with a few exceptions all the cells in a multicellular organism have the same genetic information encoded in the chains of nucleotide bases that make up their DNA. The sequence of nucleotides in the DNA chain serves as the instructions for the assembly of the proteins that engineer all cell functions. Hence it might seem that all the cells of a multicellular organism should manufacture the same proteins. In actuality, of course, each kind of cell makes only a characteristic subset of all the proteins encoded in its DNA. How is the genetic apparatus controlled so that each kind of cell manufactures a specific array of proteins and in the right amounts?

The cells of multicellular organisms are eukaryotic: they have a well-defined nucleus. In such cells the sequence of nucleotides in a gene, the stretch of DNA that codes for a single protein, is transcribed into a molecule of RNA in the nucleus. The RNA then moves out of the nucleus into the surrounding cytoplasm, where it is translated into the protein. Thus in the eukaryotic cell several levels of genetic regulation are possible. Control could be exerted over which stretches of DNA are transcribed into RNA, over which transcribed sequences are transported to the cytoplasm and over the rate at which a particular RNA is translated. It has long been assumed that the control of transcription and the control of translation play a part in the regulation of eukaryotic genes. As recently as 1975, however, knowledge of what happens to the RNA between transcription and translation was quite sketchy.

In the late 1970's it became clear that it takes several significant operations to transform the primary, or newly formed, RNA transcript into the mature messenger RNA (mRNA) needed for translation. Different structures are added to the two ends of the RNA chain. Certain nucleotides in the chain are chemically modified. Most intriguing, however, was the discovery that the transcript is frequently cut and spliced to yield an mRNA shorter than the primary transcript. In some instances the transcript can be cut and spliced in more than one way, yielding different mRNA's and hence different proteins. Such differential cutting and splicing could be a form of genetic control, and in a few instances it has been shown to be one. It is likely that posttranscription processing is necessary in many mRNA's but that it is a form of regulation in only a few. The discovery of mRNA processing has nonetheless added a new dimension to the picture of gene expression, which is in itself a significant advance. Moreover, it has given rise to intriguing hypotheses on the very early history of genes and cells.

The Transcription of DNA

Like DNA, RNA is a polymer chain whose monomer units are nucleotides. In RNA the four nucleotides are adenine, cytosine, guanine and uracil, abbreviated *A, C, G* and *U.* (In DNA the place of *U* is taken by thymine, *T.*) Each nucleotide is made up of a nitrogenous base and a five-carbon sugar. The five carbons of the sugar are generally designated by numbers. Each pair of adjacent nucleotides is linked by a phosphate group that makes a bond between the 5′ carbon on the sugar unit of one nucleo-



GENE CODING FOR A PROTEIN is a sequence of nucleotides (*A, C, G, T*) in DNA. The sequence shown here is the gene in the mouse for beta globin: one of the two chains of amino acids in the protein hemoglobin. The actual coding sequence of the DNA is interrupted by two "introns" (*gray*). The three stretches of coding sequence are "exons" (*color*). The amino acids of the beta-globin chain are shown next to the corresponding nucleotides. In order for the ge-

tide and the 3′ carbon on the sugar unit of the adjacent nucleotide. Thus the RNA molecule has directionality: one end is the 5′ end and the other is the 3′.

The transcription of DNA into RNA is accomplished by an RNA polymerase, one of three types of such enzymes in a eukaryotic cell. The enzyme binds to the DNA at a correct "start" site for RNA and selects the first nucleotide, which becomes the 5′ end of the RNA chain. It then moves rapidly down the DNA chain, adding the correct nucleotides to the RNA chain. The RNA separates from the DNA as it is assembled; the 5′ end comes off first, and the RNA molecule as a whole does not come off until the 3′ end is formed.

When the RNA polymerase copies a region of one strand of the DNA, the nucleotide that is added to the RNA is complementary to the one on the DNA. Because of the molecular structure of the nucleotides, each nucleotide can form hydrogen bonds with only one of the other three. Thus *C* and *G* can form hydrogen bonds only with each other, and *A* and *U* (or *T*) can form hydrogen bonds only with each other. It is the hydrogen bonds between complementary bases that enable two strands of DNA to form a double helix. By the same token a piece of RNA can form a double-strand molecule with a piece of DNA if the bases on the two strands are complementary.

The transcribed RNA is therefore complementary to the stretch of DNA from which it comes. Since it conserves the information from that part of the genome (the full complement of genetic information), the RNA can embody the code for the chain of amino acid units that constitutes a protein. The code is read from the RNA three nucleotides at a time, and each triplet of nucleotides, which codes for one of the 20 different amino acids, is called a codon. In the process of translation the reading of the codon results in the addition of the corresponding amino acid to the protein chain being assembled.

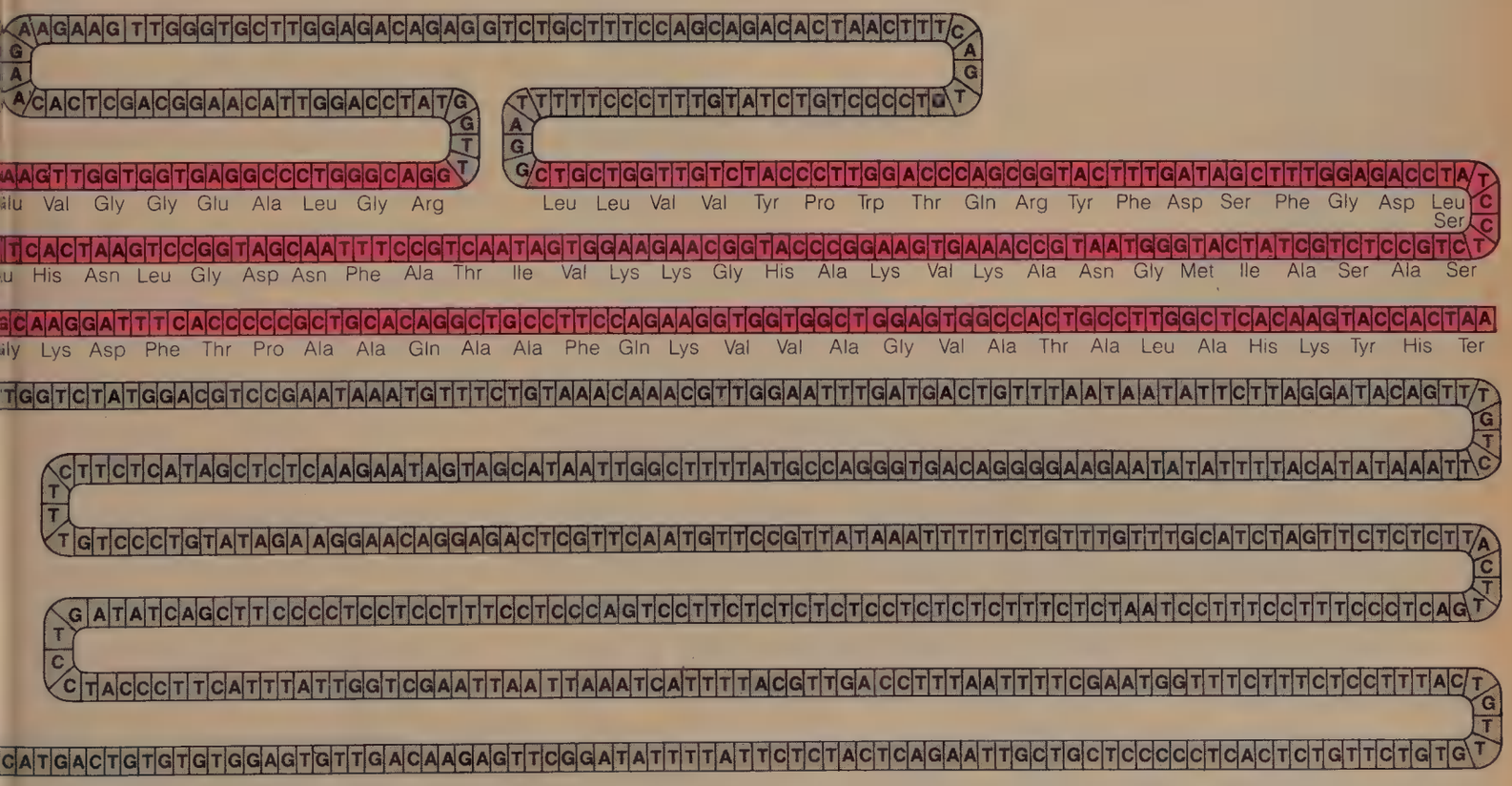
The Three RNA's

Messenger RNA, the RNA with the protein-coding function, is not the only kind of RNA. There are at least two other major kinds, each of which has a significant role in protein synthesis. Ribosomal RNA (rRNA) is a component of the ribosome, the cytoplasmic organelle that acts as a platform for translation. The ribosome is made up of two subunits, designated large and small; each subunit incorporates one rRNA molecule and between 25 and 50 protein molecules. Transfer RNA (tRNA) functions as a molecular hook to put the next amino acid on the end of a protein chain; each of the many different tRNA's recognizes a particular amino acid.

On the ribosome the transfer RNA is brought together with the messenger RNA and the enzymes whose function is to connect the right amino acid to the protein chain and to advance the apparatus to read the next codon. Thus in the synthesis of a protein, tRNA and rRNA have repetitive roles that are carried out under the instructions of the mRNA. In human cells there are about a million copies of each kind of tRNA and about five million copies of each of the two main kinds of rRNA: one for each ribosome in the cell.

Because of their abundance the transfer RNA's and the ribosomal RNA's are easier to study than the messenger RNA's, and they had been intensively investigated before much was known about mRNA. Indeed, the first example of RNA processing was found in ribosomal RNA. In 1961 Klaus Scherrer and Harriet Latham joined me in a new animal-cell and animal-virus laboratory at the Massachusetts Institute of Technology. We undertook to examine the RNA of human cells grown in a culture medium. We wanted in particular to compare the newly transcribed RNA's in the cell with the abundant kinds of stable RNA: the tRNA's and the two kinds of rRNA.

Bacteria are prokaryotic: their cells have no well-defined nucleus. In such a cell transcription and translation are carried out simultaneously: the messenger RNA begins to be translated while it is still being transcribed from the DNA. Like eukaryotic cells, however, bacterial cells rely on ribosomes and tRNA's; therefore much of the early work on the three stable kinds of RNA was done with bacterial cells.



netic message to be translated into the chain of amino acids in the protein, RNA must be transcribed from the DNA. The intron sequences must then be removed in order for the messenger RNA (mRNA) to have the correct coding sequence. Introns are characteristic of genes in eukaryotic cells: cells that unlike bacterial cells have a well-defined nucleus. The nucleotides *ATG* at the beginning of the sequence (*upper left*) are the signal for the initiation of protein synthesis. The nucleotides *TAA* at the end of the sequence are the signal for the end of protein synthesis. The nucleotide sequence was worked out by Philip Leder and his colleagues at the Harvard Medical School.

The three types of stable, abundant RNA's can be distinguished on the basis of their size. The tRNA's are all fairly small: about 80 nucleotides long. The two main kinds of ribosomal RNA are substantially larger. The size of an RNA molecule is often measured according to how fast the molecule sinks when it is spun at high speed in a centrifuge. The method is called zonal sedimentation analysis, and the sedimentation rate is expressed as the Svedberg number, abbreviated *S*. The longer the RNA, the larger its *S* is. The two major forms of rRNA in eukaryotic cells are designated 18*S* and 28*S*.

In our first experiments Scherrer, Latham and I sought to find out whether the initial products in the transcription of RNA are the same size as the abundant kinds of RNA, or alternatively whether the RNA is transcribed in units that are longer or shorter than the abundant molecules. Nucleosides, the precursors of nucleotide bases, can be labeled with various radioactive atoms. If growing cells are supplied with radioactive nucleosides, the labeled subunits will enter the pools of nucleotides that provide the building material for RNA chains. The radioactive molecules will then be added to new RNA chains by the RNA polymerases. The label provides a means of identifying the newly constructed RNA. After the labeled RNA's are extracted from the cell they can be separated according to their size by sedimentation analysis and compared with the previously existing unlabeled RNA in the cell.

The Processing of rRNA and tRNA

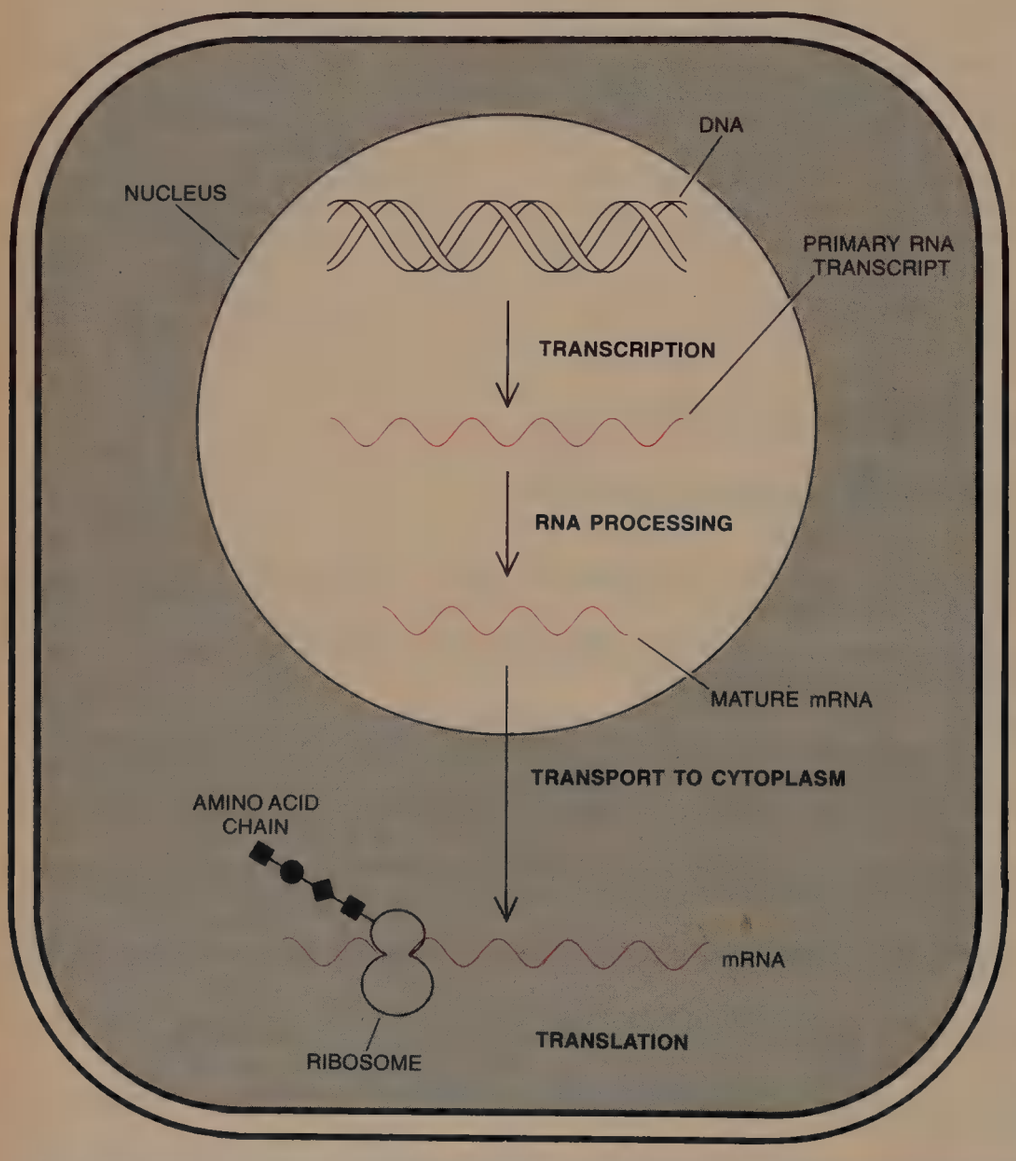
The cells were initially supplied with the labeled nucleosides for only five minutes. When this was done, the radioactivity did not appear in the RNA with sizes corresponding to the three preexisting stable kinds. Instead it appeared in the nucleus of the cells in an assortment of molecules with a wide range of sizes. Many of the labeled molecules were much longer than even the longest ribosomal RNA, which is 5,000 nucleotides long. As we shall see, such large, rapidly labeled molecules turned out to be of great significance in the work on messenger RNA. The transcription products that were labeled after five minutes were of so many different sizes that they could not be put into a few categories on the basis of their size. When the time during which the labeled nucleosides were supplied to the cell was increased from five minutes to 60 minutes, however, distinct groups of labeled molecules having different sizes became evident. One of the largest molecules, which appeared after 25 to 30 minutes of labeling, had a sedimentation value of 45*S*.

The base composition of the 45*S* RNA (the relative abundances of *A*, *C*, *G* and *U*) was very similar to that of the 18*S* and 28*S* RNA's. The 45*S* molecule was soon found to be related to the ribosomal RNA's in another and most interesting way. Transcription can be interrupted quickly by the antibiotic actinomycin, which binds tightly to DNA. We supplied actinomycin to the cells after the label had already been taken up into the 45*S* molecule. After the antibiotic had been supplied no more RNA was transcribed from DNA; hence any label was in molecules that had been transcribed earlier. Shortly after the actinomycin was supplied the label disappeared from the 45*S* molecules and reappeared in the 18*S* and 28*S* RNA's.

We hypothesized that the 45*S* RNA is a long primary transcription product that is cut to yield the two finished ribosomal RNA's. The hypothesis has since been borne out by many experiments. Improved methods of nucleotide sequencing have made it possible to determine the sequence of bases in both the genes for the ribosomal RNA and the rRNA's themselves. It is now clear that a long rRNA precursor is cut up to yield shorter finished molecules in many types of eukaryotic cells, including those of yeast, plants, frogs and mice. Recent experiments have shown that it is so even in bacteria.

Moreover, in the late 1960's Roy H. Burdon of the University of Glasgow and Deborah Bernhardt in my laboratory at M.I.T. found that transfer RNA is also formed by processing. In tRNA, however, the difference in length between the primary transcript and the finished molecule is only between 20 and 30 bases: a precursor about 100 nucleotides long is shortened to yield the 80-nucleotide finished product.

By the early 1970's the processing of ribosomal RNA and transfer RNA had been conclusively established in many



GENETIC REGULATION in a eukaryotic cell has several levels. In the nucleus of such a cell the DNA is first transcribed into a primary RNA transcript, which is a complementary copy of a stretch of DNA. The primary transcript is processed into mature mRNA: messenger RNA with the noncoding sequences removed. The mRNA is then transported from the nucleus to the cytoplasm, where it is translated by ribosomes into the amino acid chain of a protein. The expression of the genes can be controlled during transcription and processing in the nucleus, during translation in the cytoplasm and during transport from the nucleus to the cytoplasm.

types of cell. The question then arose of whether messenger RNA was processed in a similar way. It seemed unlikely that mRNA could be processed in prokaryotic cells, since in prokaryotes ribosomes translate the mRNA while it is still being transcribed. In eukaryotic cells, on the other hand, where transcription and translation are separated in time and space, mRNA processing was a distinct and interesting possibility.

RNA's in the Polyribosome

Much more had to be found out about messenger RNA, however, before the question of its processing could be approached directly. When the processing of ribosomal RNA was discovered in the 1960's, the mRNA of eukaryotic cells had not even been isolated. It had been shown quite early that amino acids are added to proteins in polyribosomes: clusters of ribosomes in the cytoplasm. It was therefore assumed that the messenger was in the polyribosomes, but it had not been proved.

In an attempt to find the putative messenger Sheldon Penman, Yechiel Becker and Scherrer in my laboratory at M.I.T. grew the human tumor cells called HeLa cells in the presence of labeled nucleosides. The nucleosides were incorporated into RNA; later labeled RNA was extracted from the polyribosomes. Some of the extracted RNA was of course ribosomal RNA, but the rRNA had not been labeled after a brief exposure to the radioactive precursors. There was, however, another group of RNA molecules in the polyribosomes that had been rapidly labeled, and these molecules had a base composition quite different from that of rRNA. The polyribosomal RNA's, as they were called, were from 500 to 3,000 nucleotides long. We hypothesized that the polyribosomal RNA included an assortment of all the messenger RNA's in the cell.

The base composition of the polyribosomal RNA's was compared with that of the long molecules that had been found in the nucleus in the course of the work on ribosomal-RNA processing. The assortment of rapidly labeled nonribosomal RNA's had been named heterogeneous nuclear RNA (hnRNA). It was readily shown that the hnRNA had a base composition much like that of the briefly labeled polyribosomal RNA. Hence it appeared that hnRNA could be a precursor of the messenger molecule that was utilized in translation in the polyribosomes.

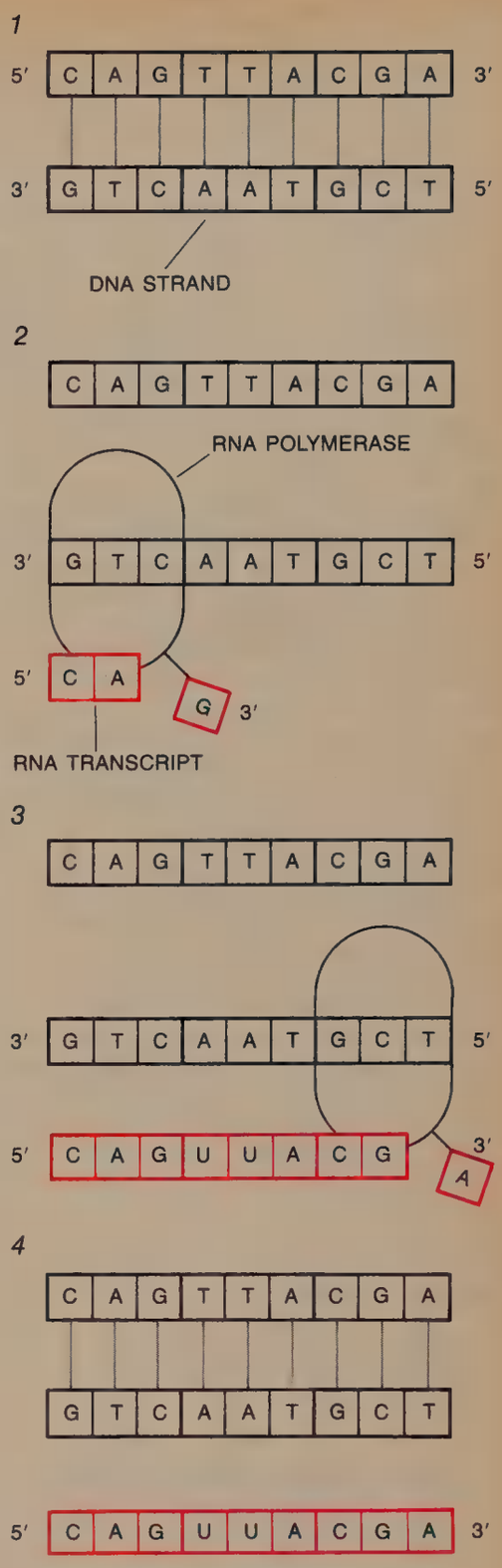
The trouble with the hypothesis was that some of the hnRNA's were 5,000 nucleotides long or longer, whereas the longest polyribosomal RNA's were only 3,000 nucleotides long. With the example of tRNA and rRNA processing in mind, we suggested that the hnRNA is shortened in a processing operation in the nucleus before being transported into the cytoplasm. It was not feasible for some years, however, to test this possibility in a direct way. The reason is that eukaryotic cells manufacture thousands of kinds of messenger RNA. Since no single mRNA had been isolated, it was not possible to observe the synthesis of an mRNA molecule to tell whether or not it is processed.

One way of circumventing the difficulty is to take advantage of the simplicity of organisms much less highly organized than cells: viruses. We therefore turned to the study of virus-infected eukaryotic cells. The infection of an animal cell by a virus turned out to be an excellent experimental system for studying the biochemical properties of messenger RNA and ultimately its synthesis. The genome of a virus, which is contained in the core of the virus particle, is a molecule of DNA or RNA. When the viral genome enters the host cell, it usurps the cell's synthetic apparatus so that the cell manufactures viral nucleic acids and proteins. At the end of the process of infection new viral molecules combine into virus particles that can infect other cells. Compared with the mammalian host cell the virus has few proteins and therefore few messenger RNA's. This simplifies matters. Since the viral mRNA must be translated by the host's ribosomes and tRNA's, the viral mRNA can be examined almost as if it were the mRNA of a eukaryotic cell. Furthermore, some DNA viruses enter the cell nucleus, where enzymes make viral mRNA; hence the steps in RNA manufacture can also be studied by means of a virus.

The Virus as a Tool

The simplicity of the virus was initially exploited to show that the briefly labeled polyribosomal RNA is the messenger. Penman and Becker in my laboratory found that in cells infected with the virus of poliomyelitis, an RNA virus, the viral RNA took the place of the briefly labeled cellular RNA in the ribosomes. Several investigators then utilized a second method to show that the viral mRNA enters the polyribosomes. The method is the one called molecular hybridization, which was first applied to DNA viruses that infect bacteria by Sol Spiegelman and Benjamin D. Hall of the University of Illinois. Molecular hybridization relies on the fact that strands of RNA and DNA can hybridize, or form a double-strand structure, only if their nucleotide sequences are complementary. If a hybrid is formed, it can be inferred that the RNA was originally transcribed from the DNA.

RNA extracted from the polyribosomes of cells infected with DNA viruses was mixed with the DNA of the virus and the host cell. When this was



TRANSCRIPTION of DNA into RNA preserves the genetic information encoded in the DNA. Each DNA nucleotide can form hydrogen bonds with only one other nucleotide: *A* (adenine) pairs with *T* (thymine) and *C* (cytosine) pairs with *G* (guanine). The nucleotides that can form hydrogen bonds are said to be complementary; the DNA helix is made up of two complementary DNA chains held together by such bonds (*1*). RNA is a single chain of nucleotides in which thymine is replaced by uracil (*U*). Both DNA and RNA chains are directional: one end is designated 5′ and the other 3′. One strand of DNA is transcribed into RNA by the enzyme RNA polymerase (*2*). The enzyme moves along one strand of the DNA, reading the nucleotides in turn and adding the complementary nucleotide to an RNA chain growing from 5′ to 3′ (*3*). When the 3′ end of the RNA is formed, the new RNA transcript separates from the DNA (*4*).

done, the polyribosomal RNA hybridized with the DNA of the infecting virus but not with the DNA of the mammalian host or that of other viruses. Brian R. McAuslan of Princeton University went on to show that the core structure of the vaccinia virus, a large DNA virus, can synthesize RNA from nucleotide precursors. Molecular hybridization was employed to demonstrate that the RNA synthesized by the virus core was the same as the polyribosomal RNA in cells infected with vaccinia virus. The polyribosomal RNA had therefore unquestionably been identified as the messenger: it was known to be present at the site of translation and it had been matched with the complementary stretches of DNA in the viral genome.

The identification of the messenger was a decisive advance. By studying viral mRNA's it was possible to follow the synthesis of specific messenger molecules to see what operations are performed on the transcript. Several significant types of messenger-RNA processing were soon found. Joseph R. Kates of the University of Colorado at Boulder noted that a segment of each mRNA synthesized by the vaccinia virus is rich in adenine nucleotides (*A*). By chemical analysis Kates showed that the adenine-rich segment is at the 3′ end of the messenger. Soon four groups led by Kates, Mary Edmonds of the University of Pittsburgh, George Brawerman of Tufts University and me independently demonstrated that almost all polyribosomal mRNA's from animal cells have a sequence of from 150 to 200 *A*'s attached to their 3′ end. The sequence is not found, however, in transfer or ribosomal RNA's.

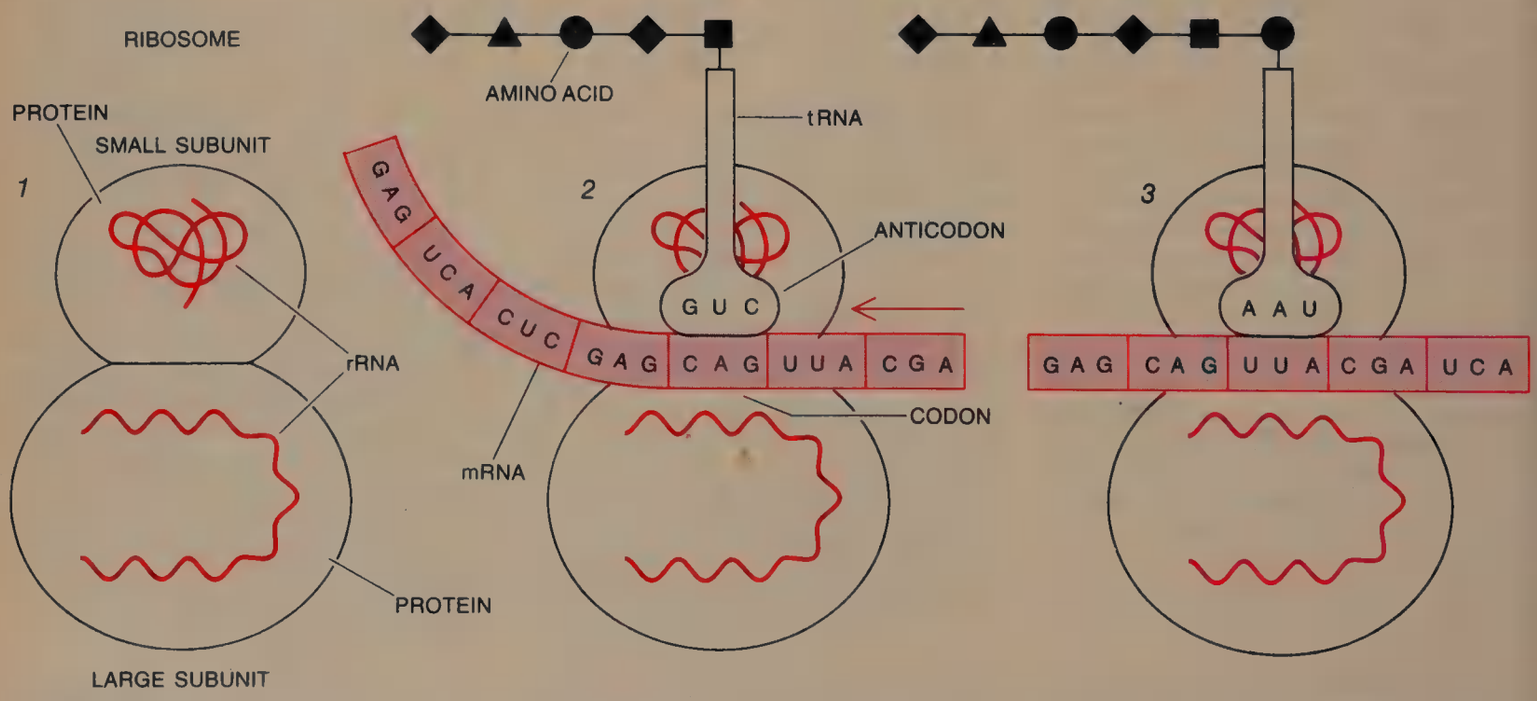
The 3′ "tail" of adenine nucleotides on the messenger RNA has been named the poly(*A*) sequence, and the process whereby poly(*A*) is added is called polyadenylation. It is now known that the tail is added to hnRNA in the cell nucleus by an enzyme almost immediately after the hnRNA is transcribed. Although the sequence of molecular events in polyadenylation has not been reconstructed in the test tube, it has now been established that the RNA polymerase proceeds far past the poly(*A*) site: the place on the DNA that corresponds to the 3′ end of the hnRNA where the tail is added. The RNA chain is cut about 20 nucleotides past a signal sequence; part of the signal is the group *AAUAAA*, which has been found in more than 100 mRNA's in plants and animals. The free 3′ end is then elongated by the addition of about 200 *A*'s. The addition is done by an enzyme discovered by Edmonds.

Tails and Caps

Almost all messenger RNA's that are transported to the cytoplasm have a poly(*A*) tail. An mRNA, however, need not have a poly(*A*) tail in order to be translated. The function of the tail appears to be to protect the mRNA from breakdown by enzymes in the cytoplasm. Messenger RNA's without the poly(*A*) tail last only a few minutes in the cytoplasm, whereas those with the tail can be stable for hours or days. The only known mRNA's that enter the cytoplasm without a poly(*A*) tail are the ones for the small proteins called histones, which bind to DNA. These histone mRNA's, designated poly(*A*)$^-$, last less than 30 minutes in the cytoplasm.

In addition to its theoretical interest the poly(*A*) sequence was of great practical benefit to the work on messenger-RNA processing. The reason is that an mRNA molecule can be caught by its poly(*A*) tail. This is done by synthesizing a polynucleotide homopolymer: a chain consisting of only one type of nucleotide. A vertical glass column is filled with paper fibers that have such a polymer adhering to them; a homopolymer consisting of all *T*'s or all *U*'s is generally used. Now the entire RNA complement of the cell is passed through the column. The poly(*A*) sequences at the end of the mRNA's hybridize with the homopolymers and remain in the column. The RNA without poly(*A*), mainly transfer RNA and ribosomal RNA, is washed away. This technique, affinity chromatography, was the first practical means of isolating in a chemically pure form the mRNA from a eukaryotic cell.

The availability of purified mRNA gave a great impetus to the work on RNA processing. Two additional processing operations were soon found. In 1974 Robert P. Perry of the Institute for Cancer Research found that animal-cell mRNA is not composed only of the four regular nucleotides. Some of the



TRANSLATION of RNA into protein requires two kinds of RNA in addition to the mRNA that bears the genetic information. Ribosomal RNA's (rRNA's) are structural components of the ribosome (*1*). Transfer RNA's (tRNA's) serve as a hook to add an amino acid to the growing protein chain; each tRNA is specific to one amino acid. Each amino acid in the protein chain is specified by a codon: a sequence of three nucleotides in the mRNA chain. The tRNA molecule binds to the mRNA by means of an anticodon, the group of three nucleotides that are complementary to the codon (*2*). As the ribosome moves along the messenger molecule the anticodon recognizes the codon and forms hydrogen bonds with it; at the same time the amino acid at the other end of the tRNA is attached to the protein chain. When the ribosome moves to the next codon, the tRNA is ejected and replaced by another tRNA that corresponds to the new codon (*3*).

nucleotides in the chain are methylated, that is, they have extra methyl groups (CH_3) added to them. Bernard Moss of the National Institute of Allergy and Infectious Diseases, Aaron J. Shatkin of the Roche Institute of Molecular Biology and Kin-Ichiro Miura of the National Institute of Genetics in Japan found that some nucleotides in viral mRNA's are also methylated.

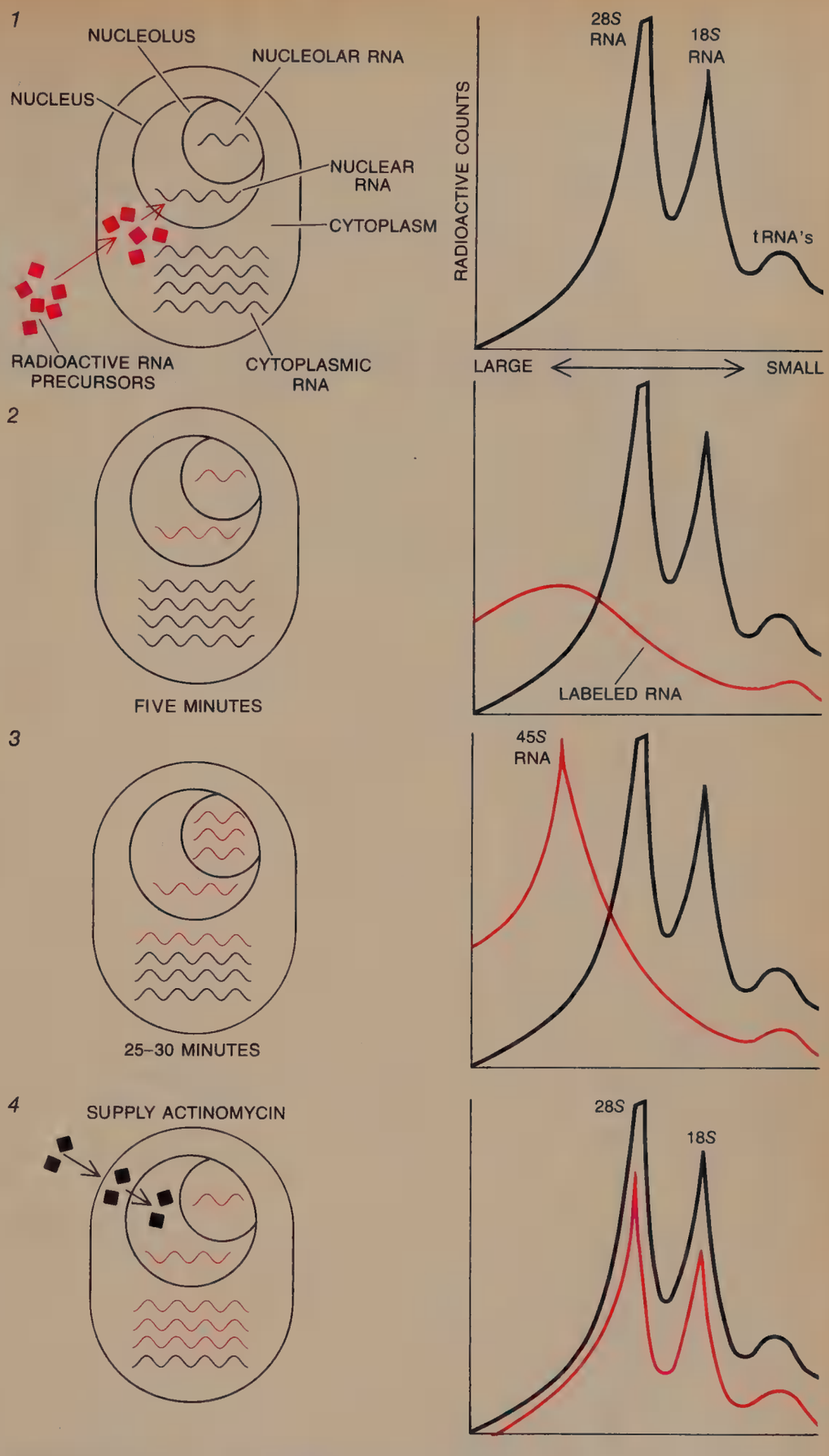
The enzymes called ribonucleases were utilized to determine where the most significant methylated structures in the messenger RNA are. Ribonucleases break the bonds between adjacent nucleotides in the RNA chain. The result is a collection of single nucleotides, each with an attached phosphate unit. When ribonucleases were applied to the methylated viral RNA, however, a complex structure that included several phosphate groups was found in addition to the single nucleotides.

Many chemical tests have been done on the ribonuclease-resistant structure, and its makeup is now understood. The structure, called a cap, is found on the 5′ end of all mRNA's that have been observed in eukaryotic cells with the exception of a single virus mRNA. The most distinctive component of the cap is a guanine nucleotide (*G*) to which a methyl group has been added. The guanine nucleotide is connected to the 5′ end of the RNA by a bond that includes three phosphate groups.

Lessons from the Adenovirus

Perry's group, Shatkin's group and my own (now at the Rockefeller University) tried to determine if the cap is added in the nucleus, as we had done when the poly(*A*) tail had been found to be a common part of cellular mRNA's. We discovered that the caps are added to the free 5′ end of the growing RNA chain by enzymes in the nucleus before the polymerase has transcribed more than about 20 bases. Moss's group has recently isolated the enzyme that adds the cap. Unlike the poly(*A*) tail, the cap is added to all mRNA transcripts. The function of the cap is not completely understood, but it appears that the structure serves to promote translation. Since the cap is added to all mRNA's, its addition probably is not a significant form of genetic regulation.

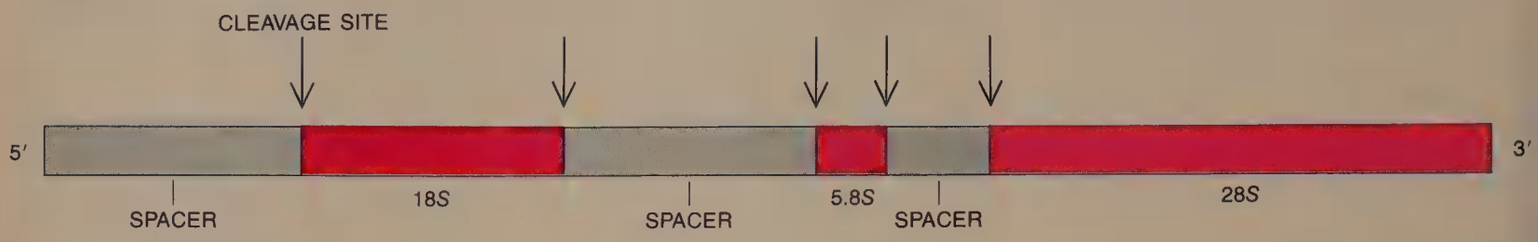
The discovery of polyadenylation and capping redirected our attention to the relation between hnRNA and mRNA. It was already known that the two kinds of molecule have a similar base composition. When hnRNA and polyribosomal mRNA were purified by affinity chromatography, it was found that the molecules in both groups have a 5′ cap and a 3′ poly(*A*) tail. Therefore if hnRNA is the precursor of mRNA, the ends of the molecule are preserved in the course of processing and transport into the cyto-



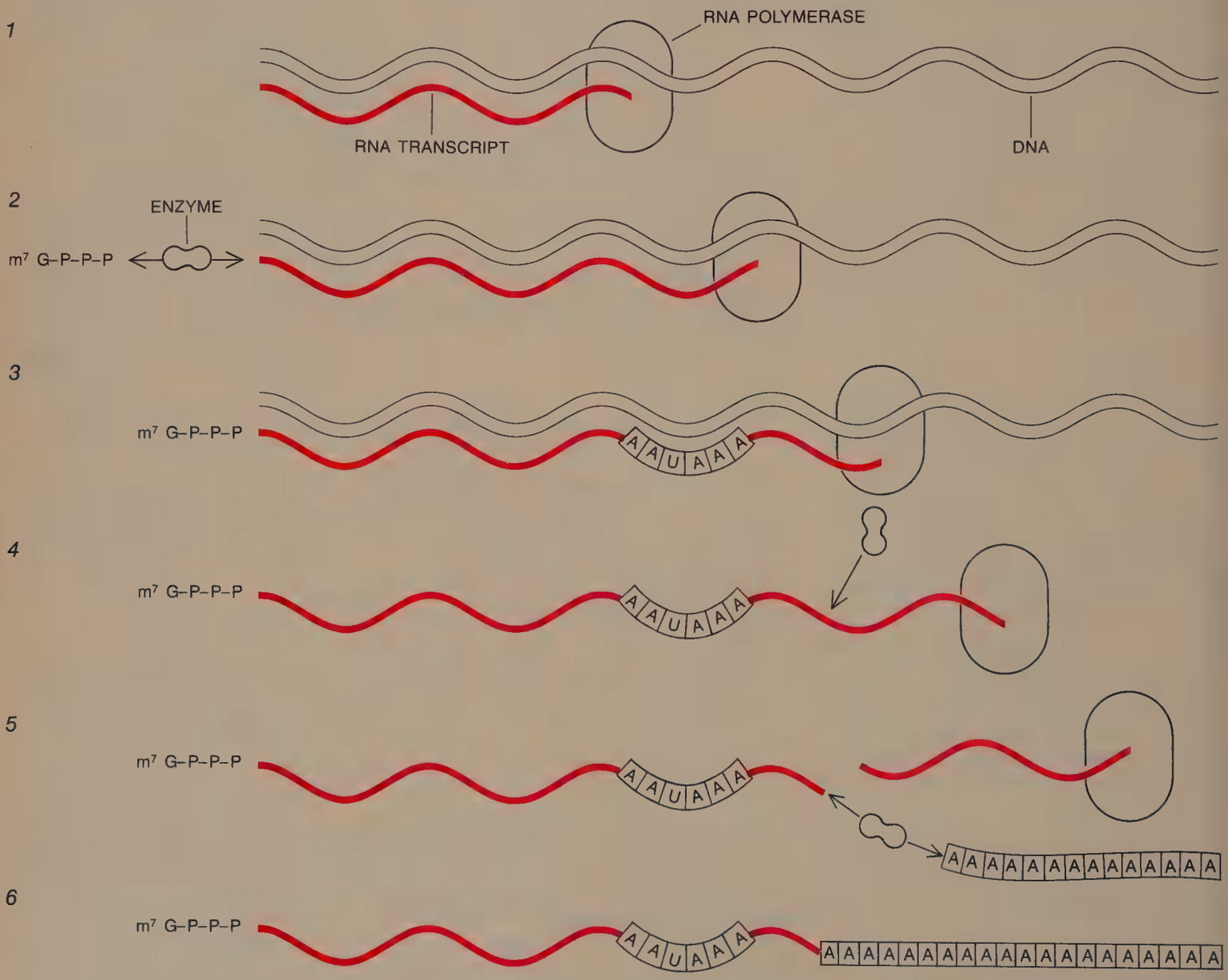
EVIDENCE FOR RNA PROCESSING was first found for rRNA. In the mammalian cell most of the stable RNA is found in the cytoplasm. A much smaller quantity of stable RNA is found in the nucleolus, an organelle of the nucleus. The size of an RNA molecule is measured by how fast it sinks when it is spun at high speed in a centrifuge; the unit of size determined in this way is *S* (for Svedberg unit). The three kinds of stable RNA are the 18*S* and 28*S* rRNA's and the smaller tRNA's. In experiments radioactive nucleotide precursors can be incorporated into new RNA, and the radioactively labeled molecules can yield information about the size of the RNA transcript. In the experiments on RNA processing mammalian cells were exposed to radioactive precursors (*1*). After five minutes the radioactive label appeared in RNA's in the nucleus with a wide range of sizes (*2*). After 25 to 30 minutes much of the label was in a nuclear molecule with a size of 45*S* (*3*). The antibiotic actinomycin, which stops RNA synthesis, was applied to the labeled cells (*4*). The label disappeared from the 45*S* molecules and appeared in 18*S* and 28*S* molecules in the cytoplasm. The 45*S* RNA therefore appeared to be a precursor that is shortened to yield the stable types of rRNA before transport to the cytoplasm.

plasm. The hnRNA's, however, were on the average five times as long as the polyribosomal mRNA's. One possible explanation for the difference seems intuitively obvious: a piece could be cut out of the middle of the hnRNA and the molecule could be spliced with the ends preserved. To molecular biologists in 1976, however, the idea appeared farfetched or even impossible.

In less than two years the farfetched possibility was shown to be what actually happens in a eukaryotic cell. Much of the work done to demonstrate that hnRNA is shortened to become mRNA utilized as part of the experimental system the adenovirus: one of many viruses that can cause infections of the upper respiratory tract in human beings. The course of adenovirus infection has been thoroughly studied. In particular the timing of the replication of the viral DNA in the cell nucleus, the composition of the viral proteins and the se-



PRECURSOR OF RIBOSOMAL RNA in eukaryotic cells includes three "spacers" that are removed in processing. Improved methods for determining the nucleotide sequences of DNA and RNA have made it possible to show precisely how the primary RNA transcript, which in human cells has a sedimentation value of 45*S*, is processed to yield the 18*S* and 28*S* rRNA's. The general structure of almost all eukaryotic primary ribosomal RNA transcripts is similar; the one shown is from the cultured human tumor cells designated HeLa. The first step in processing is for enzymes to make a cut between the 18*S* and 5.8*S* segments. The spacers are then removed by other enzymes, and the 5.8*S* RNA is attached to the 28*S* RNA by hydrogen bonds, leaving the two stable forms of RNA that are found in the ribosomes.

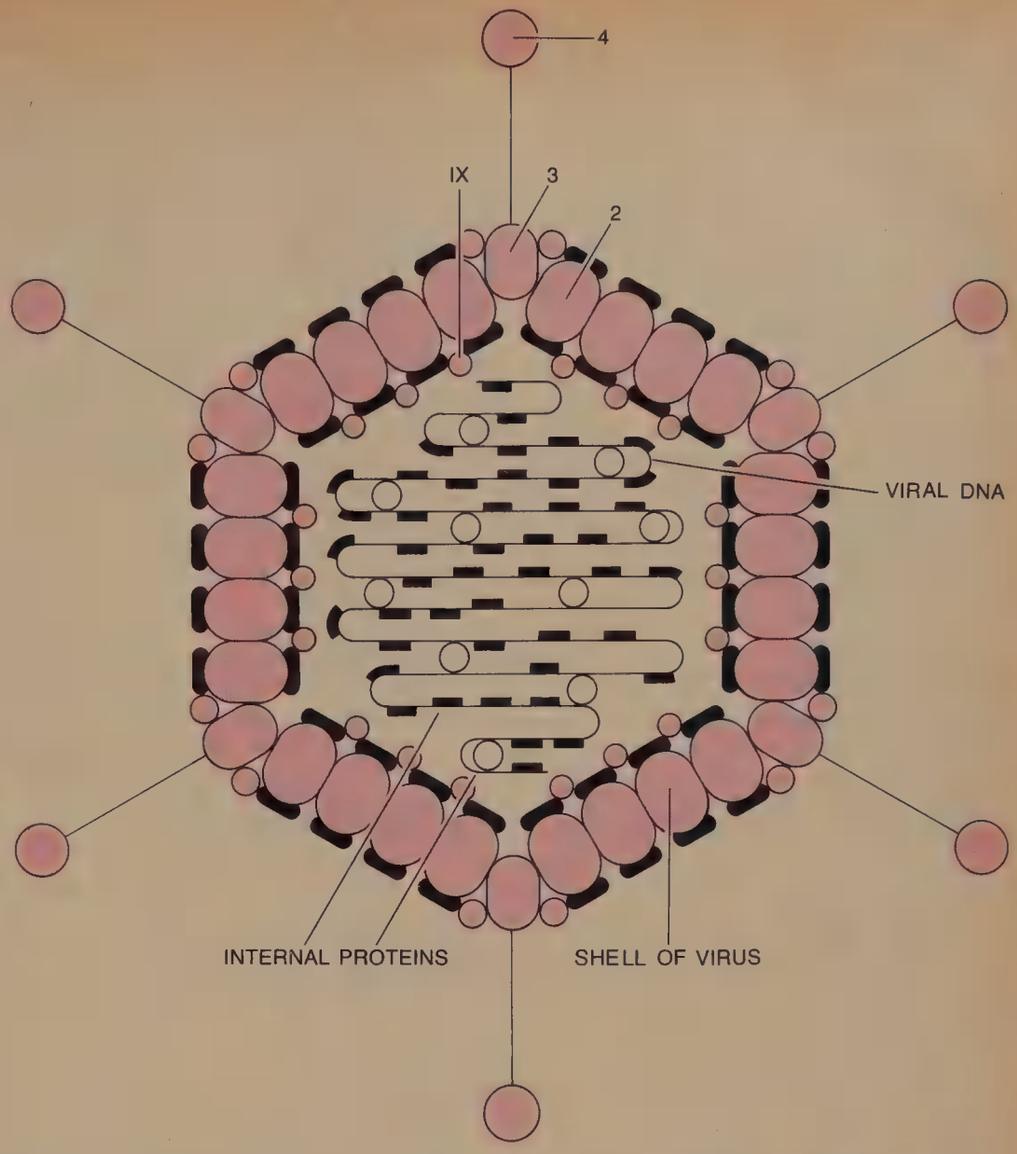


CAP AND TAIL are added to the RNA transcript in the first stages of mRNA processing. When the RNA polymerase has transcribed only about 20 nucleotides, enzymes in the nucleus add a molecular structure to the free 5′ end of the transcript (*1*, *2*). This cap consists of a modified guanine (*G*) attached to the transcript by a bond that includes three phosphate groups (*P*). As the polymerase moves farther down the DNA it transcribes the sequence *AAUAAA*, which has a role in the addition of the string of *A*'s called poly(*A*) (*3*). The RNA transcript is cut about 20 bases past the *AAUAAA* sequence (*4*). The same enzyme or another one then adds the 150 to 200 *A*'s of poly(*A*) to the 3′ end of the chain (*5*). In further processing the end structures are preserved while the interior of the transcript is modified.

quence in which the proteins are manufactured are well understood. The adenovirus was therefore a good organism to be studied for finding out how virus-specific RNA is synthesized in the cell nucleus and processed to yield mRNA.

The adenovirus genome is a double-strand molecule of DNA 36,000 nucleotides long. The position of a particular sequence of nucleotides in the viral genome is generally described according to a linear map of 100 units in which each unit stands for 360 nucleotides. Let 0 refer to the left end of the map and 100 to the right end. From eight to 10 hours after the virus particle enters a human cell the cell begins to make copies of the viral DNA. The viral proteins are classified in relation to the time of DNA replication. The "early" proteins are made before the DNA is replicated. The small group of "intermediate" proteins are made at about the time the DNA is replicated. The "late" proteins are made mainly after replication is under way. Most of the proteins that form the structure of the virus particle are late proteins. Among them are the hexon proteins, the most abundant proteins in the virus's outer shell.

The investigation of messenger-RNA processing concentrated on the late adenovirus proteins. The work relied heavily on the restriction endonucleases, a group of enzymes that had a profound effect on molecular biology when they were discovered in the 1970's. Each restriction endonuclease cuts DNA or RNA at a specific sequence of from four to six nucleotides. In a particular genome the restriction sites, the points where the enzyme cuts the DNA, are fixed. Therefore restriction endonucleases can be employed to cut a molecule of DNA into reproducible fragments. When the restriction sites have been identified, the fragments can be grouped in an ordered set with each fragment occupying a known position on the linear map of the viral genome.



ADENOVIRUS, a virus that causes upper-respiratory-tract infections in human beings, is shown in schematic cross section. This virus was an important subject in early investigations of the processing of mRNA, because at that time no mRNA's had been isolated from mammalian cells. The outer shell of the adenovirus is made up of protein molecules, most of which are those named hexons. The major hexon proteins are the ones designated 2, 3, 4 and IX. Within the shell is a double-strand molecule of DNA about 36,000 nucleotides long. When the virus enters a human cell, the viral DNA is replicated and a relatively small number of mRNA's are transcribed from it. The virus then usurps the apparatus of protein synthesis in the host cell to make proteins corresponding to these mRNA's. The hexon proteins are part of a group of "late" proteins that are synthesized mainly after the viral DNA has begun to be replicated.

Progress through Hybridization

It can readily be seen that having such a set of ordered pieces of DNA greatly increases the power of the molecular-hybridization technique. If labeled molecules of RNA are mixed with an ordered set of DNA pieces, the resulting hybrids indicate with great precision the position in the genome from which the RNA was transcribed. With this formidable technique we attempted to determine whether the adenovirus late-protein messenger RNA's are transcribed in the cell nucleus as a single long unit and then cut into smaller pieces or whether they are transcribed in units about the size of the messenger molecules.

When transcription of the viral RNA had begun, cells infected with adenovirus were exposed to radioactive nucleosides for such a short period that only a few labeled precursors were added to each RNA chain. In a population of infected cells there are RNA polymerases at almost every position on the viral DNA that serves as a template for the transcription of RNA. If label is supplied to the cells, radioactive nucleotides will be found on the ends of an assortment of RNA molecules with different lengths. The shortest labeled RNA's will be close to the starting point for transcription. Longer molecules will also be observed, with the longest near the end point of transcription.

The labeled RNA was extracted from the cells and purified. The purified RNA's were separated according to size and hybridized with an ordered set of adenovirus-DNA fragments. The molecular hybridization gave us a way of testing the hypotheses about how messenger RNA is made. If the late-protein mRNA's were transcribed in small units corresponding to individual mRNA's, the hybridization would yield a complex pattern in which short RNA's would be hybridized to the DNA near each of the multiple start sites for transcription.

Now consider what would happen if the late-protein messenger RNA's were transcribed as a single unit and then cut in pieces. There would be only one starting point for transcription, and the shortest RNA's would hybridize only with the DNA near the single starting point. Longer labeled RNA's would hybridize at the appropriate intermediate positions in the transcription unit. When the briefly labeled nuclear RNA from infected cells was hybridized to the adenovirus DNA, the shortest RNA's hybridized near only one position: map unit 16. The longer RNA's were spread out to the right of unit 16 according to their size. We therefore concluded that

the late-protein messenger RNA's are transcribed as a single unit beginning near map position 16.

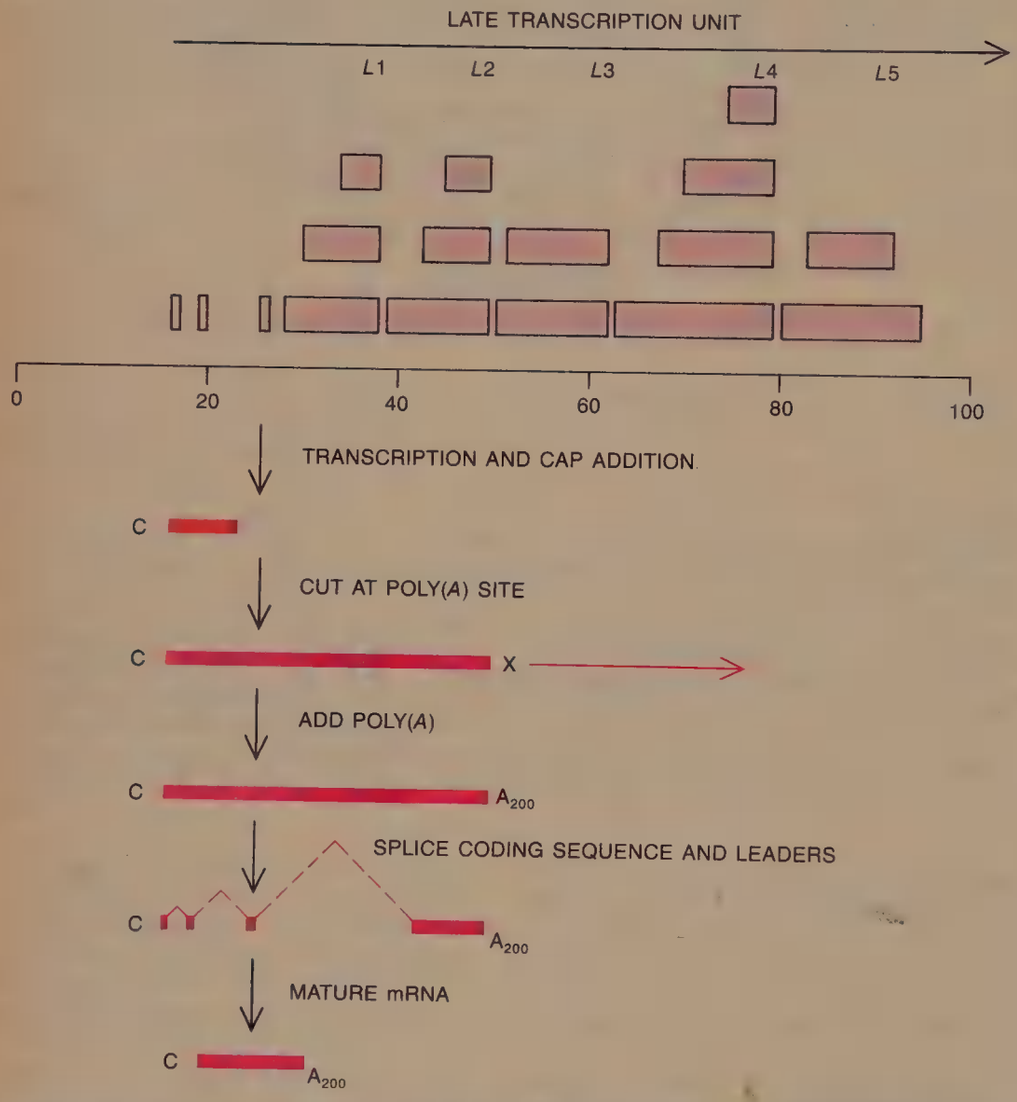
Positions on the Gene Map

Two groups of young investigators, one group at the Cold Spring Harbor Laboratory and the other at M.I.T., did a series of experiments that revealed how the primary RNA transcript from the late-protein transcription unit is processed. Richard J. Roberts and Richard E. Gelinas of Cold Spring Harbor found that all the adenovirus late-protein messenger RNA's had the same 11 nucleotides adjoining the 5′ cap. Working with the mRNA for one of the hexon proteins, Philip A. Sharp, Susan M. Berget and Claire Moore of M.I.T. found that the common sequence adjacent to the cap did not hybridize next to the major portion of the mRNA. When the hexon mRNA was mixed with DNA, it consistently formed a hybrid about 4,500 bases long in the middle of the viral genome that could be visualized by electron microscopy. Two tails, however, extended from the central double-strand structure. At the 3′ end was the poly(*A*) sequence, which was not expected to hybridize because it is not derived from the viral DNA. At the other end of the mRNA the common 5′ sequence also extended from the hybrid as a single-strand tail.

Both groups quickly recognized the possibility that the 5′ sequence might come from another part of the adenovirus genome. Sharp at M.I.T., and also Thomas R. Broker and Louise T. Chow at Cold Spring Harbor, mixed many late-protein messenger RNA's with the ordered set of DNA fragments and got an amazing result: each mRNA hybridized at four separate positions on the viral genome. Each formed a hybrid at a unique site that corresponded to the coding sequence for the protein. In addition all the mRNA's, regardless of the position of the coding sequence, formed hybrids at positions 16, 20 and 27. The result suggested that each adenovirus late-protein mRNA is a mosaic made up of three common sequences with one unique piece of genetic information.

Recent work by several investigators, including Heschel J. Raskas of Washington University School of Medicine and Joseph R. Nevins and Edward Ziff of Rockefeller University, has shown how the adenovirus mRNA's come to incorporate both common and unique sequences. In this work the adenovirus late-protein transcription unit has been shown to be a very complex piece of biological machinery. It is now known that in a mammalian cell infected by an adenovirus the main transcription product made after DNA replication is the late-protein transcription unit, which carries coding sequences for 13 or 14 late proteins. The transcription unit extends from map position 16 almost to position 100 at the right end of the map. Near unit 16 is a stretch of nucleotides called the late promoter, which induces the polymerase to begin transcription. At positions 16, 20 and 27 are the common sequences observed by Sharp, Broker and Chow. From position 27 to position 100 are the coding sequences.

The coding sequences in the late transcription unit are arranged in five groups. The groups are designated *L*1, *L*2, *L*3, *L*4 and *L*5 (in order from left to right on the map). Each numbered group includes the coding sequences for a family of related proteins. At the right end of each group is a poly(*A*) attachment site. For example, the *L*2 group extends from about position 40 to position 50; its poly(*A*) site is at position 50. Thus there is one cap site at position 16 and a total of five poly(*A*) sites in the late transcription unit.



ADENOVIRUS LATE TRANSCRIPTION UNIT has both multiple poly(*A*) sites and multiple splice sites. It can be processed to yield the mRNA's that encode more than a dozen adenovirus late proteins. A transcription unit is a continuously transcribed stretch of DNA. The adenovirus late unit extends from position 16 almost to position 100 on a linear map of the viral DNA in which each unit stands for 360 nucleotides. In all late mRNA's there are sequences complementary to positions 16, 20 and 27. From position 30 to near position 100 are the sequences that encode the late proteins, arranged in families designated *L*1, *L*2, *L*3, *L*4 and *L*5. Each family includes the coding region for several related proteins. At the right end of each family is a single poly(*A*) site common to all the mRNA's in that group. Soon after transcription begins a cap structure is added to the 5′ end; the RNA polymerase eventually transcribes the entire late unit. As it traverses the five poly(*A*) sites one of them is chosen; other enzymes cut away the nucleotides at that site and poly(*A*) is added. Here *L*2 has been chosen. The sequences between positions 16, 20 and 27 and between position 27 and the *L*2 family are removed. One of the coding segments in the *L*2 family is then chosen. The basis for the choice of poly(*A*) site and splice site is not known. The mature mRNA includes the 5′ cap, the common sequences at positions 16, 20 and 27, the code for a particular late protein and the 3′ tail.

Splicing the Messenger

For each family of proteins the coding sequences have a similar overlapping arrangement. All the coding sequences overlap at the right end of the numbered unit, and all of them include the nucleotides adjacent to the poly(*A*) site. In the middle of the group, however, the coding sequences diverge: each one includes a different segment of the nucleotides in the middle of the unit. When a mature messenger-RNA molecule is

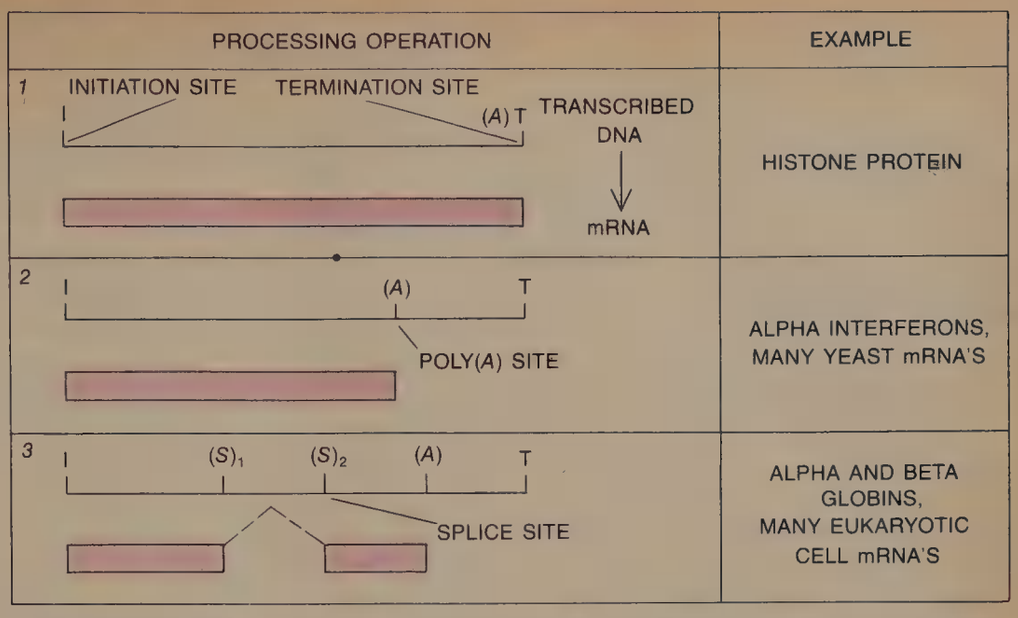
formed from the late transcription unit, only one coding sequence is included.

From many experiments it has become clear that the mature mRNA is processed in the following way. Transcription is initiated at the late promoter and moves to the right from position 16. When a few nucleotides have been transcribed, the dangling 5′ end of the RNA is capped. The polymerase then moves toward position 100, passing all five poly(*A*) sites. During the passage enzymes cut the transcript at one poly(*A*) site and attach the poly(*A*) tail there. The first RNA molecule in the stages of processing therefore includes all the sequences from position 16 to the chosen poly(*A*) site. For example, if the *L*2 site is picked, the RNA will include the sequences from position 16 to position 50; the *L*1 family of coding sequences is included with the *L*2 family but not the *L*3, *L*4 and *L*5 families. The RNA downstream from the *L*2 site (including the *L*3, *L*4 and *L*5 groups) is apparently destroyed in the nucleus and is never utilized in a messenger molecule.

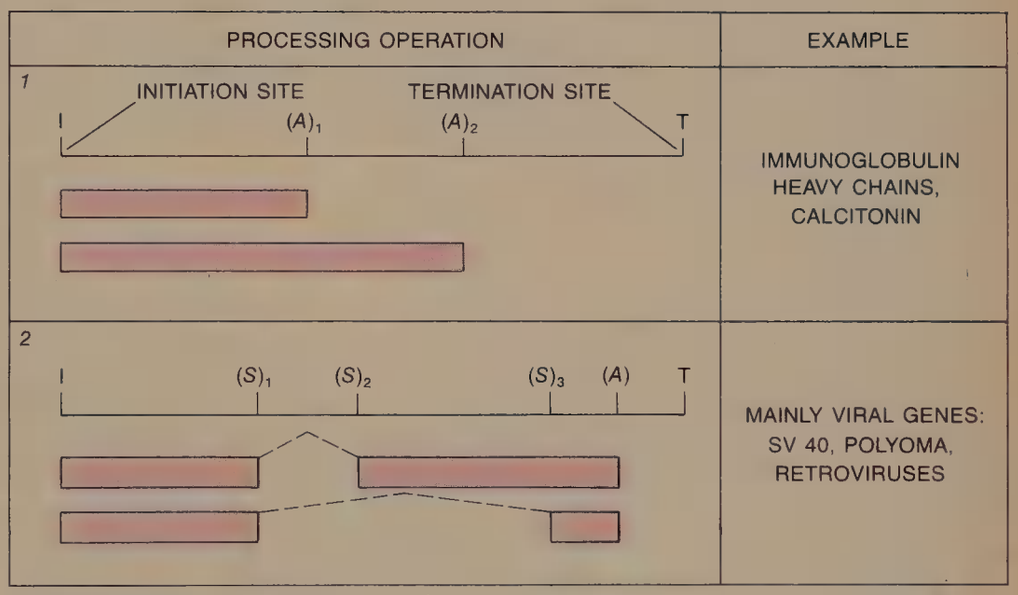
The partially processed poly(*A*)-containing nuclear RNA therefore includes several "spacers" that are not part of the mature mRNA. The spacers are removed by enzymes in a process that can take as long as 20 to 30 minutes. The *L*1 sequences are taken out. The stretches from position 16 to position 20 and from position 20 to position 27 are also removed, leaving only the common sequences at positions 16, 20 and 27. The result is a molecule with one group of related coding sequences. In this instance *L*2 has been selected. The *L*2 family has three coding sequences. One of the three is picked by enzymes that remove a variable piece of RNA from within the *L*2 family. The result is the mature mRNA, with the 5′ cap, the common 5′ sequences, the poly(*A*) tail and one unique coding sequence.

The nucleotides from position 16 to position 100 in the adenovirus genome make up an extreme example of a complex transcription unit. Such a unit is one that can be processed to make more than one kind of mature messenger. A simple transcription unit, on the other hand, yields only one kind of mRNA. By a combination of multiple poly(*A*) sites and differential splicing a formidable array of late proteins can be processed from a single adenovirus transcription unit.

It can readily be seen that either multiple poly(*A*) sites or differential splicing alone could yield a complex transcription unit. Only a few complex transcription units have been found in eukaryotic cells. All but one of the complex eukaryotic units have multiple poly(*A*) sites only. Only one has both multiple poly(*A*) sites and differential splicing: the unit for the heavy chain of the protein myosin in the fruit fly.



SIMPLE TRANSCRIPTION UNIT can be processed to yield only one kind of mRNA. The processing of simple units is shown in schematic form at the left; examples of such units and the organisms in which they are found are given at the right. The initiation site is where transcription begins; the termination site is where it ends. The poly(*A*) site corresponds to the point where the 3′ poly(*A*) tail is attached. In some instances the RNA polymerase stops at the poly(*A*) site (*1*). In other instances transcription proceeds past the poly(*A*) site (*2*). The transcript can be cut and spliced to yield a shorter finished mRNA if there are noncoding sequences within it (*3*). The majority of transcription units found in mammalian cells are simple ones.



COMPLEX TRANSCRIPTION UNIT can be processed to yield more than one mRNA and hence more than one protein. The presence of two or more poly(*A*) sites can result in a complex unit (*1*). If the unit is cut at the first poly(*A*) site, one mRNA is made; if it is cut at another, a different mRNA is made. In a transcription unit with only one poly(*A*) site there can be more than one splice site. The mRNA that is made depends on which splice sites are chosen (*2*). Complex units with more than one poly(*A*) site are common in viruses and have also been found in mammalian cells. Units with differential splicing have been found in viruses and one example has been found in a eukaryotic cell. Complex virus transcription units with multiple splice sites can be integrated into the genes of the host cell and expressed along with the host's genes.

In the genome of the rat a single transcription unit with two poly(*A*) sites has been found by Ronald Evans, Michael Rosenfeld and their colleagues at the Salk Institute for Biological Studies. The unit includes the coding sequence for two proteins: the hormone calcitonin, synthesized in the thyroid, and a newly discovered neuropeptide, a nervous-system hormone synthesized in the pituitary. The two proteins have identical "heads" (the end of the protein chain designated the amino terminal) but different "tails" (the end designated the carboxy terminal). In the cells of the thyroid gland the first poly(*A*) site is picked and calcitonin is made. In the cells of the pituitary gland the second poly(*A*) site is picked and the neuropeptide is made.

The work done since 1975 has proved that genes in eukaryotic cells can be regulated by means of RNA processing. It is not yet clear, however, how significant such processing will turn out to be as a form of genetic regulation. For most eu-

Amateur Telescope Making

Edited by Albert G. Ingalls
Foreword by Harlow Shapley

This three-book set is the authoritative reference library of the enthralling hobby of amateur telescope making. Through these books thousands have discovered a fascinating mechanical art combined with a great science.

BOOK ONE begins at the beginning, teaches the basics of glass grinding and how to complete the first telescope. (510 pages, 300 illustrations.)

BOOK TWO leads on into advanced methods of amateur optical work and describes new projects for the telescope maker. (650 pages, 361 illustrations.)

BOOK THREE opens up further fields of enterprise: binoculars, camera lenses, spectrographs, Schmidt optics, ray tracing (made easy). (646 pages, 320 illustrations.)

SCIENTIFIC AMERICAN ATM Dept.,
415 Madison Avenue, New York, N. Y. 10017

Please send me postpaid the following AMATEUR TELESCOPE MAKING books.
My remittance of $_____ is enclosed.

- ☐ **BOOK ONE** $10.00
- ☐ **BOOK TWO** $12.50
- ☐ **BOOK THREE** $12.50

For U.S. shipments add $1.00 each; elsewhere add $2.00 each.

Name_____

Address_____

City_____

State_____ Zip_____

Residents of New York City please add city sales tax. Other NYS residents please add state sales tax plus local tax.

karyotic genes the chief method of regulation still appears to be the control of transcription. Moreover, the majority of eukaryotic transcription units that have been examined are simple units, which can encode only one protein.

Exons and Introns

Even in simple transcription units, however, processing is done to remove spacers. After the work with the adenovirus broke the ground it became clear that in eukaryotic genes the expressed sequences (exons) are also commonly interrupted by nucleotide sequences that are not translated (introns). Divided genes have been found in eukaryotes as unrelated as yeasts and human beings. In all the cells splicing is needed in order to get a messenger RNA without the introns and with the exons united for translation. Thus RNA processing is quite widespread among eukaryotes.

The divided genes of eukaryotes represent a remarkable departure from the genes of prokaryotes. In a prokaryotic cell the genes are uninterrupted, and extra pieces of DNA are soon discarded from the genome. The view of cellular evolution that once prevailed held that eukaryotes evolved from single-celled organisms much like current prokaryotic cells. If this view is correct, it follows that introns were introduced into prokaryotic genes in the process of evolution. On the other hand, Carl R. Woese and his colleagues at the University of Illinois at Urbana-Champaign have proposed that eukaryotes did not descend from prokaryotes and that there are two equally ancient prokaryotic lineages. Woese argues that the three lineages descended from a single precursor that he calls a progenote (meaning "before a cell").

The detailed analysis of the introns in eukaryotic genes supports the notion that eukaryotes did not evolve from prokaryotes. First, the intervening sequences in some eukaryotic genes have been in place for a very long time. For example, in the oxygen-carrying molecule hemoglobin there are two protein chains: alpha globin and beta globin. The amino acid sequences of the two chains differ, but the chains are folded to form similar three-dimensional structures in the hemoglobin molecule. The structures are also quite similar to the structure of myoglobin, the oxygen-binding protein in muscle cells.

In spite of these structural similarities, the differences in the amino acid sequences of alpha globin, beta globin and myoglobin suggest that the genes for the three proteins diverged from a common precursor perhaps a billion years ago. At the same time all three genes have three exons separated by two introns in analogous positions in the three genes. Where the amino acid sequences diverged, the arrangement of introns appears to have been preserved, so that the introns could have been in place for a billion years.

Moreover, when the three-dimensional structure of a protein is examined, it is often found that the protein chain can be divided into segments having different functions; such segments are termed functional domains. Many introns divide genes into pieces each of which codes for a different functional domain. Therefore if the introns were introduced into an existing genome at random, they had the remarkable effect of dividing genes into pieces corresponding to separate functions. It seems more likely that the introns have been in place perhaps since the beginning of evolution and that the exons coding for functional domains evolved independently of one another.

In the Beginning Was RNA

On the basis of these arguments and others W. Ford Doolittle of Dalhousie University in Nova Scotia, Darryl C. Reanney of Latrobe University in Australia and I have proposed that even in precellular times the information for proteins was not continuous. It is widely thought that even though DNA comes before RNA in the sequence leading to the synthesis of protein in the cells of today, the first coding nucleic acid was not DNA but RNA. The reasons include the fact that primitive RNA chains can be synthesized without enzymes. Furthermore, RNA can store coding information and is necessary in translation. If RNA evolved as a random nucleotide chain, its useful information might not have been continuous. RNA splicing to bring discontinuous pieces of information together would thus have been an early and propitious evolutionary event. Recently Thomas R. Cech of the University of Colorado at Boulder has shown that RNA can be spliced without enzymes, at least in the protozoan *Tetrahymena*. Therefore it is not implausible that intervening sequences existed from the earliest times and that RNA splicing was an early feature of gene expression.

Such evolutionary speculations will not be resolved for some time. Meanwhile pressing detailed questions remain to be answered. An intensive effort is under way to identify the enzymes that splice the transcript. When they are found, it will be possible to discover how one splice is selected rather than another. Similarly it must be learned why one poly(*A*) site is selected rather than another. On a more general level is the question of how significant RNA processing is in genetic regulation as a whole. Whatever the answers may be, the investigation of RNA processing will continue to have a profound effect on our understanding of how genes are expressed in eukaryotic cells.

BREAKTHROUGH: FLOAT AN EYE TO STEADY ITS GAZE.

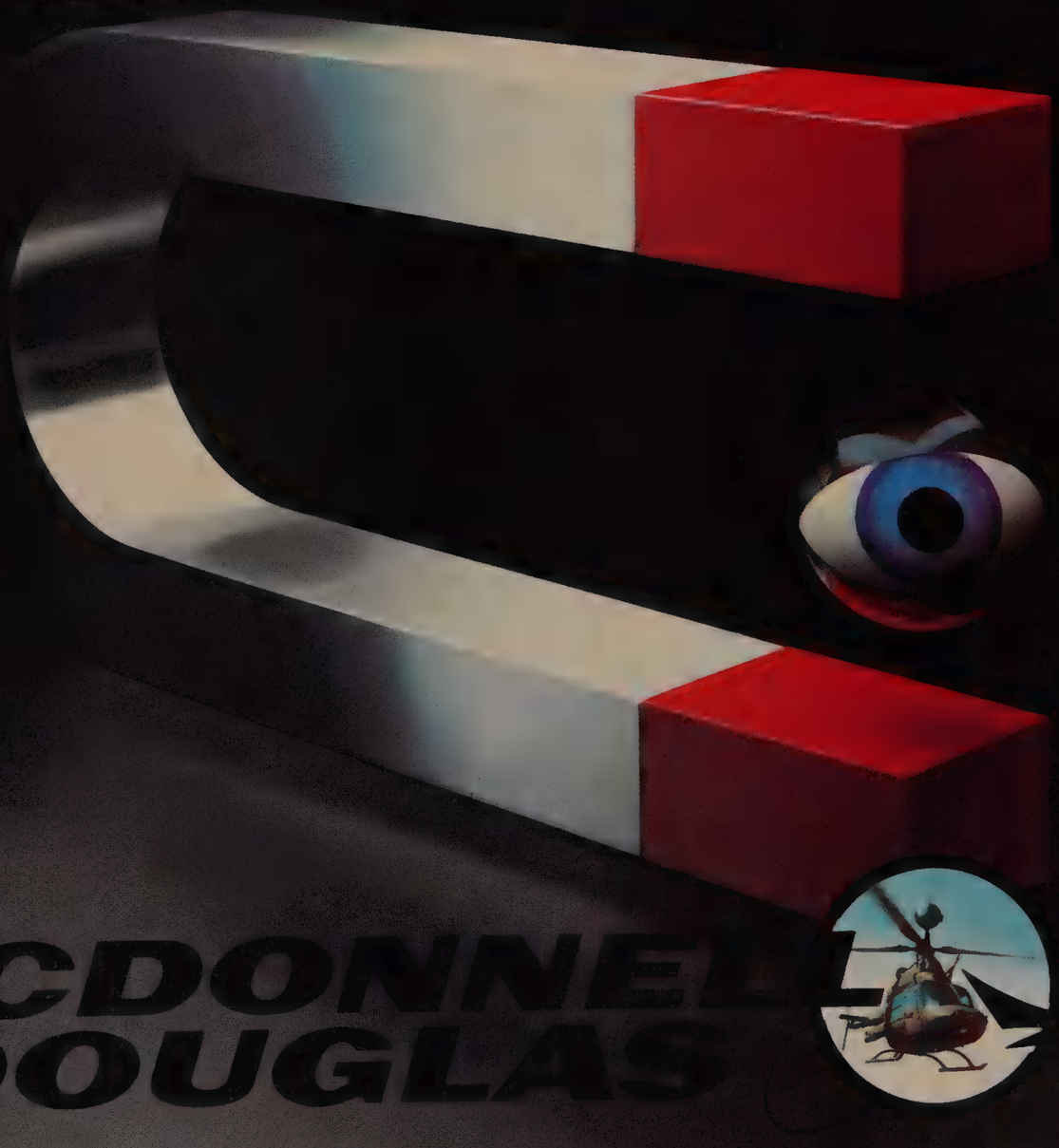
No matter how well protected mobile cameras or telescopes may be, the motion of the vehicle carrying them causes the lenses to jitter and shake. Images are blurred and imprecise.

We discovered a way to make a moving lens sit still. By using high-energy, rare earth magnets, we suspend an electro-optical system in a magnetic field—in effect, make it float. Untouched, it remains steady while its source of transportation vibrates. This "soft ride" assures crisp, unblurred vision for an observer or a sensor.

Army scouts call it the first major advance in mobile observation since binoculars were issued to the cavalry.

We're creating breakthroughs not only in optics but also in aerospace, health care, energy and information processing.

We're McDonnell Douglas.



MCDONNELL DOUGLAS

The Dead Sea

Five years ago the world's most hypersaline lake "turned over," dissipating a dense body of fossil brine that had been isolated from any contact with the atmosphere by fresher water above it

by Ilana Steinhorn and Joel R. Gat

For thousands of years fresh water, carrying its light burden of salts leached from rock, sand and soil, has flowed south through the Jordan Rift Valley into the Dead Sea, a terminal lake. There is no outflow. Under cloudless skies the inflowing water evaporates into the desiccated desert air, leaving the salts to accumulate in the lake and its sediments. As a result the Dead Sea is the world's saltiest natural lake, with an average salinity of 280 grams per kilogram compared with the ocean's average of 35 grams. The Dead Sea is also the lowest place on the face of the earth: its surface is about 400 meters below sea level. Both the lake's salinity and its low location make it a valuable natural resource. As a storehouse of chemicals it is already being exploited by Israel and Jordan, and Israel has plans to realize the lake's potential as a sink for the generation of electric power by piping seawater to it from the Mediterranean.

Any lake is a sensitive monitor of environmental changes, the record of which is encoded in the sediments of the lake floor. In the case of the Dead Sea geologic and hydrologic history is recorded not only in sediments but also in the briny water itself and its long-term salinity structure. Among all the hypersaline lakes of the world the Dead Sea is the deepest as well as the saltiest, reaching a maximum depth of 320 meters. For centuries the fresh water flowing onto its surface mixed to only a small extent with the much saltier lake water, tending to form less saline layers floating over a dense column of fossil water. Then, beginning in the 1960's, the diversion of increasing amounts of the fresh water for irrigation lowered the lake and made its surface water saltier and hence denser. A few years ago the age-old gradient disappeared and the water column turned over, the fossil brine mixing with the younger water. In this article we shall briefly describe the Dead Sea's geologic, hydrographic and chemical history and its current and potential economic development before telling in some detail what we and other limnologists and geochemists have been able to learn about this historic overturn.

The Dead Sea has two basins: a larger, very deep one to the north and a smaller one, only a few meters deep, to the south. The two basins are separated by a broad peninsula called El Lisan (Arabic for "the tongue") that projects from the eastern shore, leaving only a narrow strait along the western bank; the lake is so low now that the sill of the strait is exposed and the southern basin is dry most of the year (except for parts maintained as evaporation ponds). The southern end of the lake is at about the latitude of transition from a Mediterranean climate to an arid one, and indeed the Judean Mountains bordering it to the west and the Moab Mountains to the east have as much as 600 millimeters (about 24 inches) of rain per year.

Yet the local climate of the Dead Sea is extremely arid and hot. The lake is in the "rain shadow" of the Judean Mountains, and the mean annual rainfall is only between 50 and 75 millimeters. At Sedom in the south there are 300 cloudless days per year, the summertime relative humidity is from 30 to 40 percent and the average monthly temperature ranges from 16 to 34 degrees Celsius (61 to 93 degrees Fahrenheit); 54 degrees C. (129 degrees F.) has been recorded during the day in summer. Considering the aridity, water evaporates from the Dead Sea relatively slowly because the vapor pressure over the surface is lowered by the water's dissolved salts.

In the 1930's (before major irrigation works were begun) the Jordan River, which combines the outflow of Lake Kinneret (the Sea of Galilee) in the north with that of the Yarmuk River drainage basin to the east, supplied some two-thirds of the Dead Sea's total inflow. The balance was accounted for by smaller rivers such as the Arnon, by a few freshwater and saline springs and by the seasonal wadis that drain the mountainous flanks of the lake. As is typical in a zone of climatic transition, fluctuations in precipitation and runoff were rather large, and the depth and area of the Dead Sea varied accordingly.

In the past two decades most years have seen a steady decrease in inflow as, beginning in the early 1960's, much of the freshwater flow was diverted, first into Israel's National Water Carrier system and then also into Jordan's Ghor Canal. Currently the level of the lake is being lowered by about half a meter per year (except in an occasional very wet period). Even if the present rate of loss by evaporation were maintained, however, it would take many hundreds of years for the lake to dry up because the northern basin is so deep. Actually the rate of loss would decrease because the surface area would be reduced as the level falls and because the salinity would increase with the reduction of volume, further decreasing the vapor pressure and hence the evaporation rate per unit area; eventually an increasingly concentrated brine becomes hygroscopic and absorbs water from the air.

The origin of the Dead Sea can be traced in terms of plate tectonics: the movement of the great plates into which the lithosphere, the rigid outermost shell of the earth, is divided. The lake lies on the Syrio-African Rift, a "transform boundary" along which two plates have been sliding past each other for some 10 million years. The crust to the east of the rift has moved north with respect to the crust to the west. The relative motion has stretched and thus attenuated the crust, which has subsided. In places the fault zone zigzags to the west, across the direction of the sliding motion, which has opened up discrete basins such as the Dead Sea's.

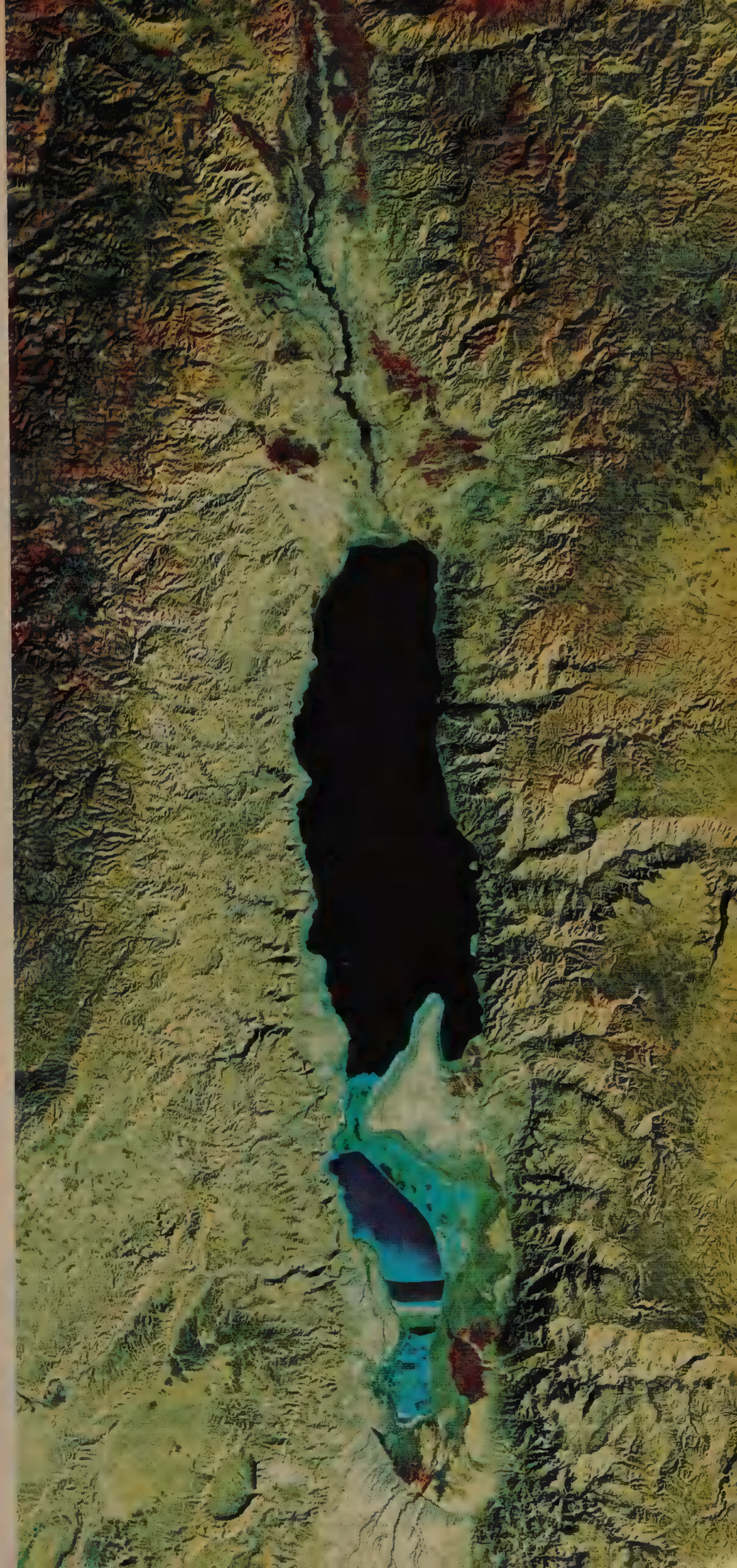
Immense layers of salt (sodium chloride) and gypsum on the Rift Valley floor reveal that during the Pliocene epoch (some two to five million years ago) the valley was in effect a great lagoon connected at times to the ocean to the west. The marine connection was disrupted by tectonic movements one or two million years ago. Thereafter a succession of lakes occupied the valley. The

best documented, Lake Lisan, was larger than the Dead Sea, much deeper (its surface was only some 180 meters below sea level) and less saline. Perhaps 10,000 years ago the valley and the Dead Sea attained what is essentially their present shape, but the level of the lake continued to fluctuate slightly.

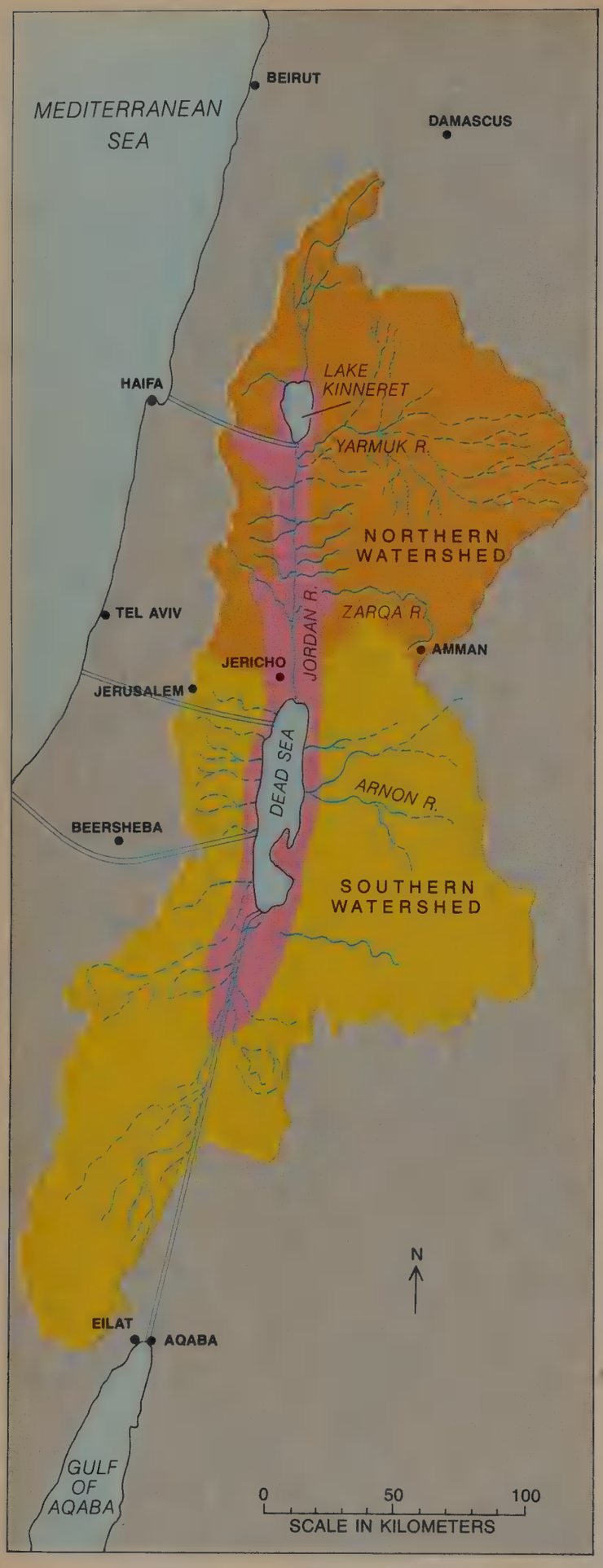
Allusions in the Bible suggest that the water was quite low, leaving the southern basin dry. The earliest map showing the Dead Sea, a mosaic from A.D. 560, shows only the northern basin. A fairly reliable reconstruction of changes in water level is possible from the beginning of the 19th century, based on floodmarks, the dating of driftwood or submerged tree trunks and travelers' reports (telling, for example, whether or not the Lisan ford 402 meters below sea level could be traversed). Actual measurements, recorded since the late 1920's, show a high-water mark of 392 meters below sea level in 1930; since then there has been a gradual decline.

The chemical character of the Dead Sea brine evolved along with the lake and its basin. The brine holds a unique assemblage of salts. Compared with brines having a marine origin, it is enriched in calcium, magnesium, potassium and bromine and is depleted in sodium, sulfate and carbonate. Such a composition could not have evolved from evaporating seawater, nor could it result simply from the weathering of continental rocks or the dissolution of minerals formed earlier by evaporation. The detailed steps in the evolution of the Dead Sea brine are still in dispute, but there must have been several sources.

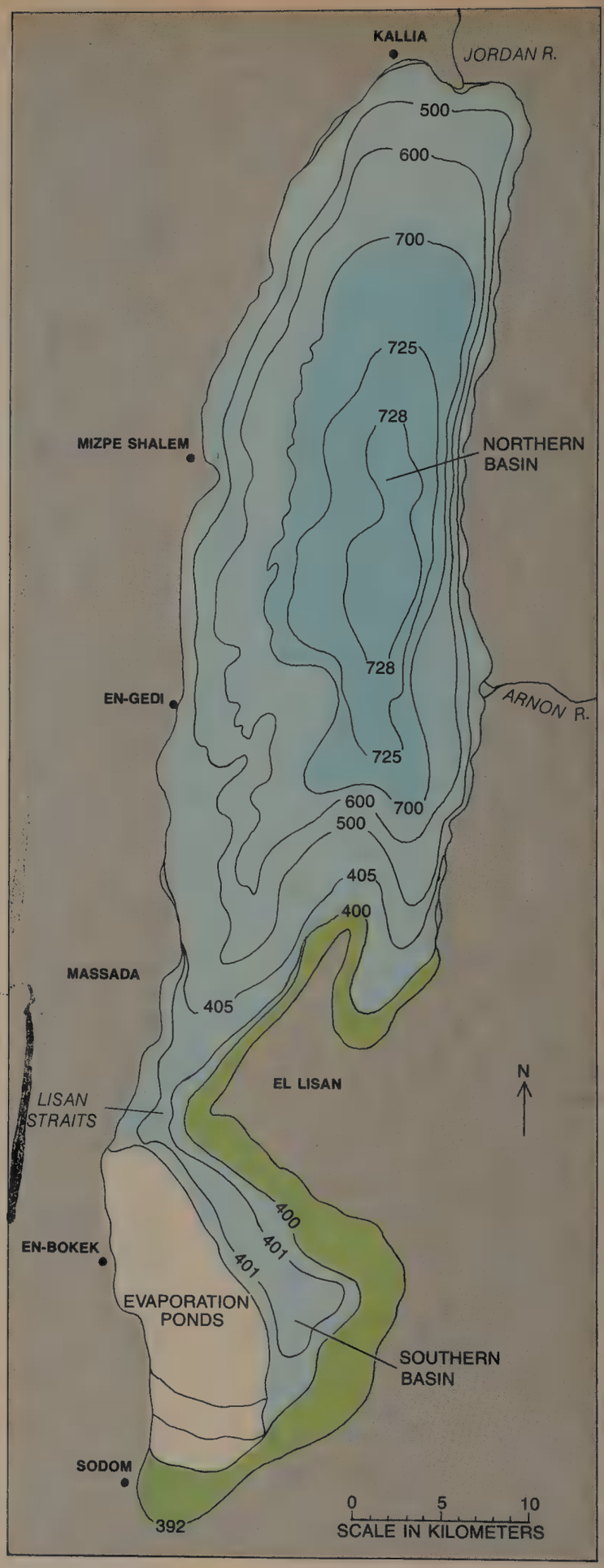
Efforts to determine the age of the brine lead to a similar conclusion. As long ago as 1715 the English astronomer Edmund Halley suggested that one might compare the rate of influx of salts with a chemical inventory of the salts in the lake and its sediments, and so learn how long it had taken for the brine to accumulate (and thus, he proposed, establish the date of the Creation). In 1961 Yaacov K. Bentor of the Hebrew University did such a calculation with data on the chemical load of the Jordan River and of known saline springs and seepages. He arrived at an age of some 12,000 years for the Dead Sea's salt as-



JORDAN VALLEY AND DEAD SEA are seen from an altitude of 570 kilometers in this false-color Landsat image. The deep saline water of the lake's large northern basin absorbs light rather than reflecting it, and so it is black. The swampy flats of the smaller, dried-up southern basin are dark green and light blue; the diked evaporation ponds are various shades of blue. Vegetation is red and urban areas are blue gray. The satellite image was acquired in December, 1978, about when the Dead Sea's waters were turning over.



NATURAL DRAINAGE BASINS of the Dead Sea are shown in yellow and orange. Most of the northern watershed's supply has been diverted for irrigation, reducing total inflow to the lake. The lowered lake's potential as a sink for the generation of electric power could be realized by bringing water to it from the Mediterranean Sea or the Gulf of Aqaba. Four proposed routes are shown; the one passing south of Beersheba is currently favored. The area below Mediterranean Sea level is in light red. Broken lines indicate seasonal wadis.



DEAD SEA'S TWO BASINS are very different, as is made clear by this bathymetric chart. Depths are indicated in meters below mean sea level. The lake as it was when the surface was 392 meters below sea level, in the 1930's, is shown in blue and green. By 1975 the shoreline had receded to about the 400-meter contour; water from the deep northern basin still crossed over the sill of the Lisan Straits into the shallow southern basin. Now the surface is even lower and the southern basin is largely dry except for the diked evaporation ponds.

semblage. Such a result implies that the salts in the lake are almost as young as the lake itself, and hence that the entire salt load of the Dead Sea's precursor lakes is sealed off under its sediments.

Closer examination reveals, however, that most of the salts are actually brought into the lake not by the major river system but by local springs and seepages. Many of the latter sources have a chemical composition very much like that of the lake itself: they are recycling the brine of the Dead Sea or that of its precursors. The accumulation of salts in the lake therefore results from a lower rate of influx than Bentor calculated, maintained for a longer time. The brine's unique composition reflects its derivation not only from the recent and current inflow of the products of continental erosion but also from the recycling of older deposits, including those of the ancient marine lagoon. Such an explanation is consistent with the Dead Sea's location at the deepest spot on the earth's terrestrial surface, from which the flushing away of accumulated salts is quite impossible.

Hypersaline water is not hospitable to life, and the absence of higher life forms long ago gave the Dead Sea its name. (The less judgmental Hebrew name is *Yam Hamelach:* the Salt Sea.) It was only in 1936 that Benjamin Elazari-Volcani of the Sieff Institute in Israel established that the lake was inhabited by microorganisms. He found several species of bacteria and one species of alga. The alga, the green *Dunaliella,* is ordinarily found in fresh water or seawater but can adapt to a hypersaline environment by synthesizing large amounts of viscous glycerol, which increases the osmotic pressure inside the cell and thus prevents the entry of salt.

Halotolerant bacteria, which are able to live in water with a moderate salt concentration, were found only at the northern end of the lake where the fresh Jordan water mixes with the lake water. A number of truly haloresistant species were more widely distributed. The most abundant species, however, were found to be aerobic, halo-obligate ones: bacteria that require oxygen and that cannot survive in media containing less than 150 grams of sodium chloride per kilogram. The halobacterial cell contains a very high concentration of potassium, and the cell membrane has an extraordinary specificity for potassium as opposed to sodium. The bacterium can therefore retain its potassium and so maintain its integrity.

There is a strong correlation between the physical and chemical properties of the Dead Sea and the number and types of microorganisms it harbors. One study done when the lake was stratified showed that the total number of bacteria decreased by a factor of 100 at a depth of 50 meters. No living algae at all were found below 100 meters. In 1970 investigators counted 40,000 *Dunaliella* cells per milliliter at the surface; below 50 meters they could find only anaerobic bacteria.

The commercially important minerals in the Dead Sea brine are potash, or potassium chloride (KCl), which constitutes 1 percent of the brine; common salt, or sodium chloride (NaCl), 8 percent; magnesium chloride ($MgCl_2$), 13 percent, and the bromide ion (Br^-), .3 percent. The first practical steps toward their extraction were taken in the 1930's by M. A. Novomeyski, a mining engineer from Siberia who received a concession for potash production from the British mandate administration. Novomeyski's process of selective precipitation is still the basis of production by the Dead Sea Potash Works. Lake water is transferred into a series of shallow evaporation pans, and each of the various salts precipitates as its concentration reaches the saturation point. Gypsum (calcium sulfate) crystallizes first, followed by sodium chloride. The mineral carnallite (a hydrated potassium and magnesium chloride) begins to crystallize when the brine reaches a specific gravity of 1.3 grams per cubic centimeter. To extract potash, whose major application is as a fertilizer, a slurry of carnallite is broken down into its components and the potassium chloride is further refined. A small fraction of the remaining "end brine" is treated with chlorine gas to extract bromine.

At first the evaporation ponds were on the shore at the southern end of the lake, but their total area was extended to about 130 square kilometers in 1966 by dikes built out into one half of the shallow southern basin. Changes in the lake since the ponds were built have affected the operation. The increased salinity shortens the time required for evaporation, but it also means increased precipitation of salt in conduits and pumps. The lowered level of the lake means that the brine has to be pumped from farther away and raised a greater distance.

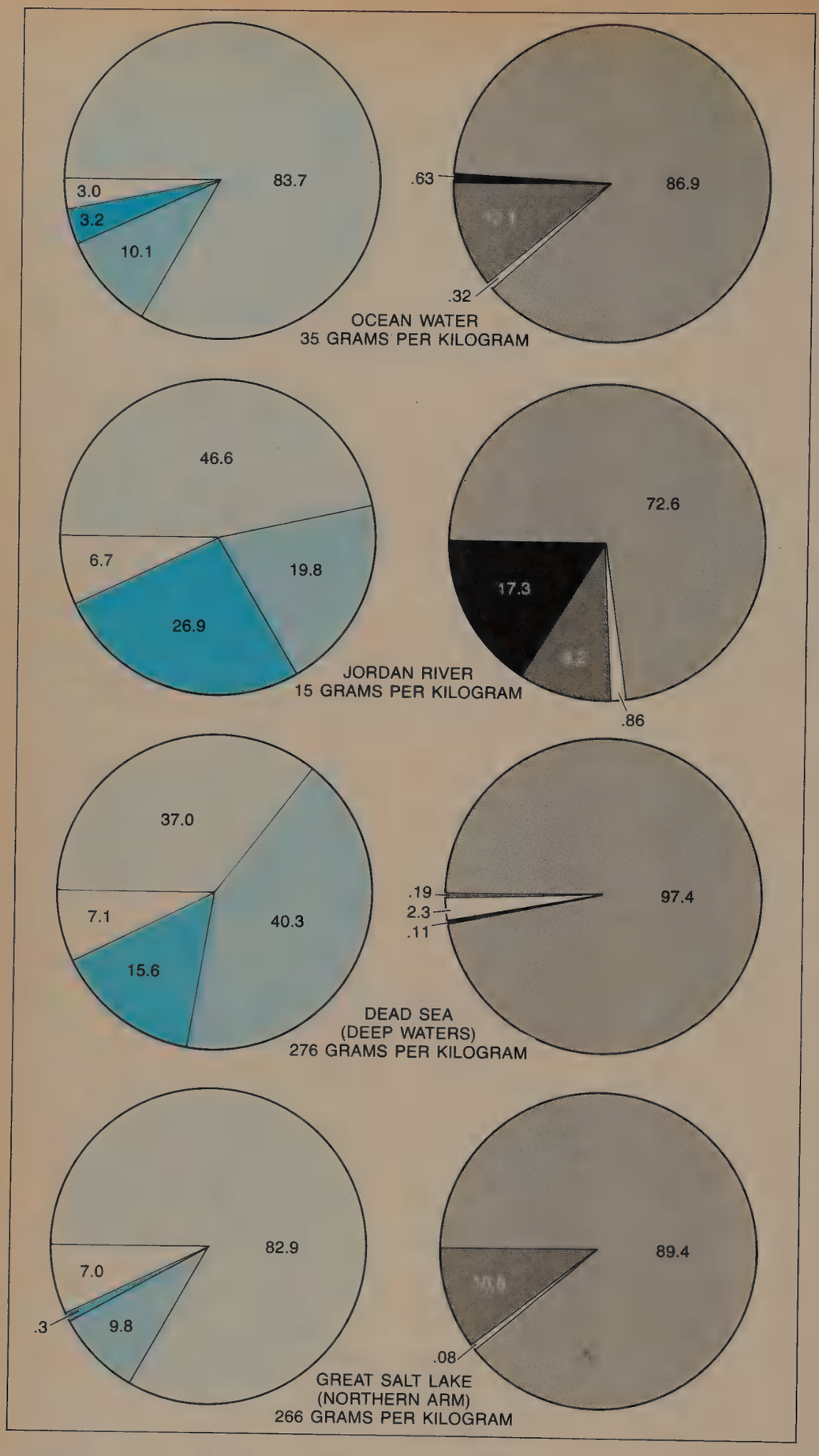
In 1902 Theodor Herzl, the visionary of the Zionist enterprise in Palestine, suggested in his book *Altneuland* that it should be possible to take advantage of the 400-meter difference between the level of the Mediterranean Sea and that of the Dead Sea to generate hydroelectric power. A broad Jordan Valley plan put forward 40 years ago by the American conservationist Walter C. Lowdermilk included a proposal to divert Jordan water for irrigation and build a canal from the Mediterranean to compensate for the loss of inflow to the Dead Sea and produce 100 megawatts of electric power. Now the Jordan has indeed been largely diverted (and the Yarmuk as well). The resulting lowering of the Dead Sea's level makes the hydroelectric scheme particularly attractive now. It makes possible a large throughput of water without any flooding of tourist and industrial sites along Israel's Dead Sea shores, at least until the lake has been refilled to its pre-1960's level.

In 1977 a planning group appointed by the Israeli Government considered four alternative routes: one from the Gulf of Aqaba to the south and three from the Mediterranean. The northernmost route (the one envisaged in the Lowdermilk plan) would put a canal through the Valley of Jezreel from near Haifa to just south of Lake Kinneret, flooding the lower Jordan Valley; in effect it would copy the ancient natural connection between the sea and the Rift Valley. This route would be expensive and would have a major environmental impact, in particular on important aquifers: permeable strata that are sources of fresh water. The other three routes would require substantial tunneling. The study group favored the southernmost route from the Mediterranean. It is not the shortest one, but it avoids the country's major aquifers and has the potential of promoting development in the northern Negev.

The plan calls for refilling the lake to a typical level of the 1930's, 393 meters below the Mediterranean, over a period of from 10 to 20 years. An inflow of about 1.6 cubic kilometers of water per year could be accommodated, which would generate up to about 800 million kilowatt-hours of electricity per year, saving some 275,000 metric tons of liquid fuel. Storage reservoirs would be built on the steep escarpment at Massada, some 300 meters above the lake so that the generation of electricity could be regulated to meet peak demand.

Once the Dead Sea has been refilled the rate of inflow will have to be modulated to the mean water-balance deficit. Even if the remaining flow of the Jordan and local flood flows are completely diverted to irrigation, the mean rate of inflow of Mediterranean seawater that will balance evaporation is estimated to be only some two-thirds as much as the rate of inflow during the filling stage, so that the hydroelectric system would inevitably become less profitable than it had been during the first 10 to 20 years. The permissible rate of inflow can be improved to the extent that the rate of evaporation from the lake's surface can be increased. The seawater's vapor pressure will be higher than that of the more saline Dead Sea brine, and so it will be helpful to avoid mixing the seawater deeply with the lake water, thus maintaining a sharply stratified layer of seawater mixed with some brine on top of the denser lake water.

The layering effect is also central to a novel "solar pond" scheme for energy



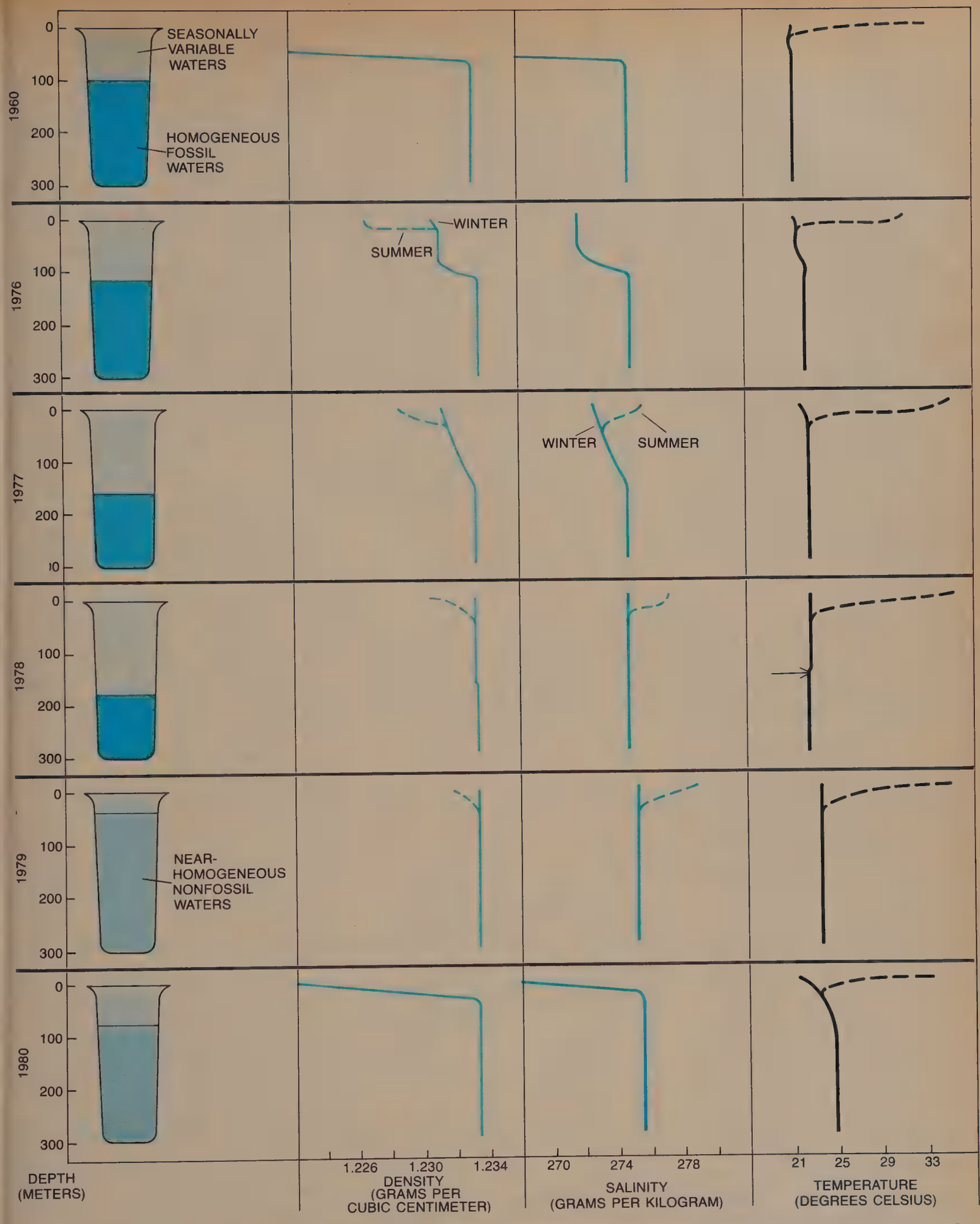
UNIQUE CHEMICAL COMPOSITION of the Dead Sea is compared with that of other bodies of water. The overall salinity of each is given in grams per kilogram. The pie charts show the percentage (on a weight basis) that is accounted for by each positively charged ion (cations) and negatively charged ion (anions). It is clear that the Dead Sea's salt assemblage could not have been derived simply from either ocean or Jordan River water.

production. If a thin layer of relatively fresh water is maintained over a briny pond, the sun heats the brine, whose density prevents it from rising and losing its heat to the atmosphere. The heat can be extracted by a heat exchanger that generates steam, and the steam powers a generator. In small-scale experimental solar ponds built near the Dead Sea, temperatures as high as 110 degrees C. (less than the boiling point of the brine) have been achieved. The mean electricity output of such a solar collector could be as much as five megawatts per square kilometer of Dead Sea area. Many physical and technological problems of floating the pools on the surface of the Dead Sea remain unresolved, however, and so the feasibility of the scheme is still not established.

Any lake in a climate such as the Dead Sea's undergoes a seasonal cycle. In the case of a freshwater lake, the sun heats the surface water in the summer, decreasing its density, while the deeper water remains cooler and denser. A stabilizing density gradient is established based on the temperature difference. A thermocline, a boundary region where the gradient is steep, isolates the largely undisturbed deep water from the top 10 meters or so that are warmed by the sun and stirred by the wind. Total circulation, with mixture of the surface water and deep water, is reestablished when the surface water cools enough in the autumn or winter. In the Dead Sea salinity as well as temperature affects the density of the water. A complex pattern of stratification ensues that can change but does not go through a simple repetitive annual cycle.

Variations in salinity arise from dilution of the brine by freshwater inflow or enrichment of the brine through evaporation. Theoretically the changes in density caused by such fluctuations in salinity could be compensated for by temperature changes in the right direction. For example, an increase of 20 degrees C. in temperature, if it were balanced by an increase of 9.2 grams per kilogram in salinity, would leave the density unchanged. In actuality, however, in this century the Dead Sea's surface salinity has varied by as much as 40 or 50 grams per kilogram, and there was even greater variation in the past. The large changes in salinity resulting from excesses of freshwater inflow over evaporation, or from the opposite condition, cannot be compensated for by the annual temperature cycle of from about 18 degrees C. in the winter to 35 degrees in the summer. The water-temperature change is usually not of sufficient magnitude to balance the change in salinity, nor are the two kinds of change likely to be precisely in phase.

The result is that the structure of the Dead Sea's water column and its season-



EVOLUTION OF WATER COLUMN (*left*) **and the density, salinity and temperature profiles that influenced it are traced from 1960 to 1980. In 1960, as for many years before, a large density difference (caused by a large change in salinity) separated the variable upper layers from the deep fossil waters. The density and salinity profiles are almost congruent on the scale chosen here (a scale at which the low surface values go off the chart). By 1976 the interface between the surface and the fossil waters had descended. The upper layers' salinity and density had increased, but in the summer the surface density was decreased by warming. These trends continued in 1977, but there was a new reverse salinity gradient in the summer. In 1978 the salinity was uniform throughout the column; a small temperature difference (*arrow*) maintained the stability of the fossil water body. In 1979, after the overturn, the fossil water body had been dissipated and the entire column was more saline and somewhat warmer. In 1980, after a rainy winter, there was a new but shallow salinity gradient.**

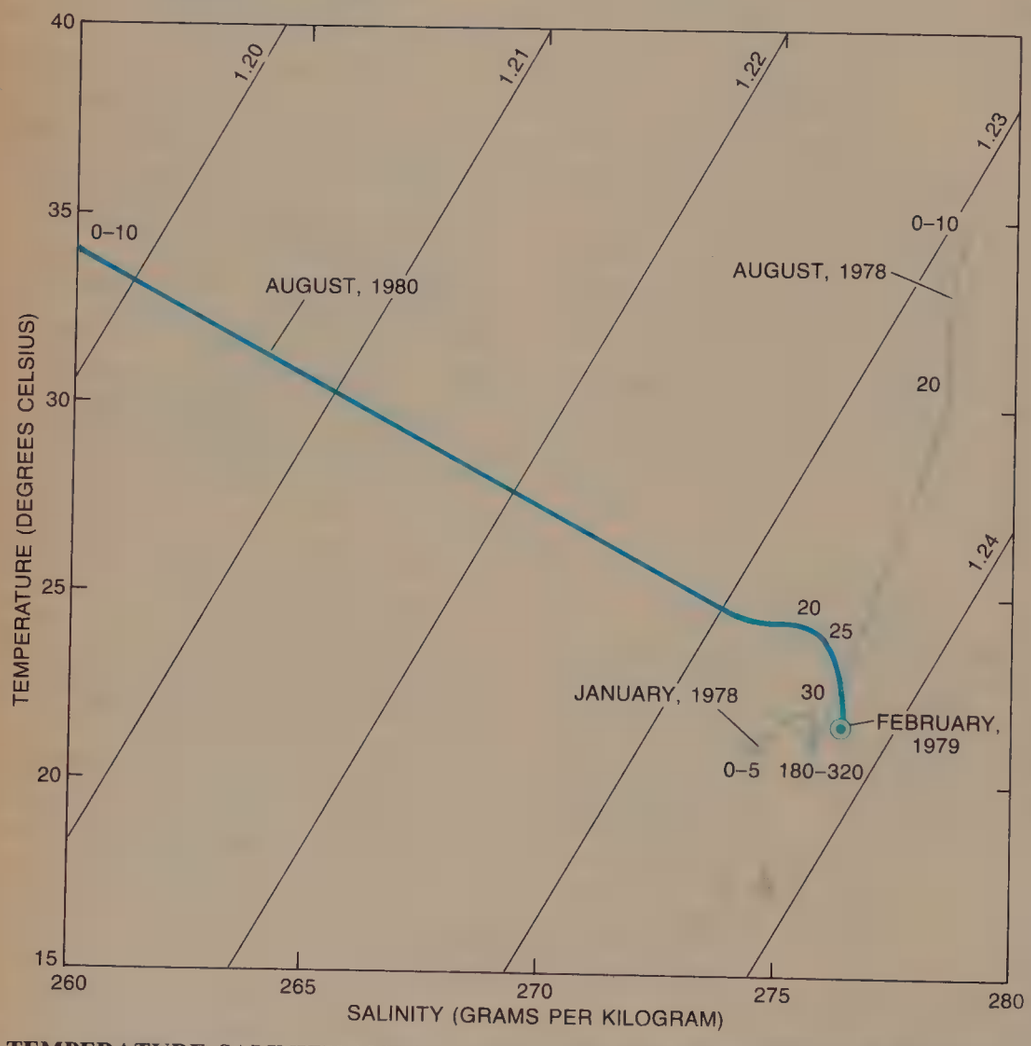
al and long-term evolution are very different from those of a freshwater lake and also (because the Dead Sea is so deep) from those of most saline and hypersaline lakes. The first hydrographic measurements, made in 1864, showed that the Dead Sea's water column was stratified by virtue of a salinity gradient. This was confirmed by measurements made in 1919; it appeared to be a long-term characteristic. David Neev of the Geological Survey of Israel and Kenneth O. Emery of the Woods Hole Oceanographic Institution studied the stratification in 1959 and 1960. They reported a density of between 1.20 and 1.21 grams per cubic centimeter (corresponding to a mean salinity of about 250 grams per kilogram) in the upper layers of the lake. A salinity gradient—a sharply defined difference in salinity of 25 grams per kilogram at a depth of from 35 to 40 meters and a more gradual transition zone going down to 80 meters—separated these upper, seasonally variable layers from the deeper water. Within the upper layers there were sharp differences in temperature and salinity that varied with the season.

Below 80 meters the water was essentially homogeneous, with a temperature of 21.3 degrees C. and a salinity of 276 grams per kilogram. That was a higher salinity than had been reported for any surface water in either basin. Moreover, a strong odor of hydrogen sulfide emanated from deep-water samples, suggesting that the water contained no dissolved oxygen and harbored anaerobic bacteria deriving their energy from the reduction of sulfates. The lack of oxygen was confirmed by the fact that iron appeared in the reduced, bivalent state. It appeared that the homogeneous water below 80 meters constituted a body of fossil water that had remained undisturbed for a long time, isolated by the density gradient from any contact with the upper layers and the atmosphere.

The fossil water body was finally to dissipate during the winter of 1978–79 by mixing with the overlying layers. The dramatic overturn and the events leading up to it were documented by hydrographic surveys carried out since 1975 by our group at the Weizmann Institute of Science in Israel. From the beginning of the survey it was clear that the salinity of the surface layers had increased significantly: the difference in salinity between the surface water and the deep water had diminished from about 25 grams per kilogram in 1959 to only two grams per kilogram in 1975; the stability of the water column was diminished accordingly. The water column was becoming more homogeneous as the mean salinity of the younger, upper layers approached that of the deep fossil water. The interface between the younger and the fossil waters was successively lowered. From a depth of between 60 and 80 meters in 1960 it descended to between 80 and 100 meters in 1975 and to 200 meters in the fall of 1978. The fossil water body was being eroded, but what remained of it stayed essentially unchanged.

By the summer of 1978 the upper layers had actually become saltier than the deep water. Yet stratification was maintained, now by virtue of a temperature gradient. The surface water was warmer, and therefore continued for a time to be just a little less dense, than the bottom water. During the next winter the nonfossil water cooled somewhat, and that was enough to bring about the final mixing. When the overturn took place, the surface water was more saline, and still somewhat warmer, than the fossil water. As a result the homogenized water column was half a degree warmer at the time of the overturn than the fossil water mass had been, and also .4 gram per kilogram saltier.

How long had the long-term stratification persisted before the overturn? How old, in other words, was the fossil water body? Marianna Stiller of the Weizmann Institute and Y. C. Chung of the Scripps Institution of Oceanography determined the age of two radioactive isotopes recovered in 1978 from below the deepened interface: radium 226 and carbon 14. They assumed that both isotopes had been introduced into the surface layers and had been mixed through the water column before its stratification. (Radium 226 originates primarily in warm saline springs along the Dead Sea's shores; carbon 14, which is present in trace amounts in atmospheric carbon dioxide, becomes dissolved in surface water that is in contact with the air.) Once the deeper water was disconnected from the surface the supply of the two isotopes could not have been replenished. The isotopes were subject to radioactive decay, with a half-life of 1,260 years for the radium and of 5,600 years for the carbon. Measurements of the extent to which they had actually decayed indicated that the water below 180 meters had begun to be isolated about 300



TEMPERATURE, SALINITY AND DENSITY of water at various depths are shown at four stages before and after the lake's overturn. The chart is a temperature-salinity diagram. A given temperature and salinity combine to yield a particular density. The diagonals are isopycnic curves: curves of equal density (in grams per cubic centimeter). The short profile of January, 1978, shows conditions were not very different at the surface from what they were at depth. By August the surface water had become more saline than the deep water, but it was warmer and so it remained less dense. In February, 1979, during the overturn, temperature and salinity were about the same throughout the water column, so that the entire profile is contracted to a single point. In August, 1980, a large inflow of less saline waters had temporarily reestablished a density gradient, and there was a shallow, sharp pycnocline, or density boundary.

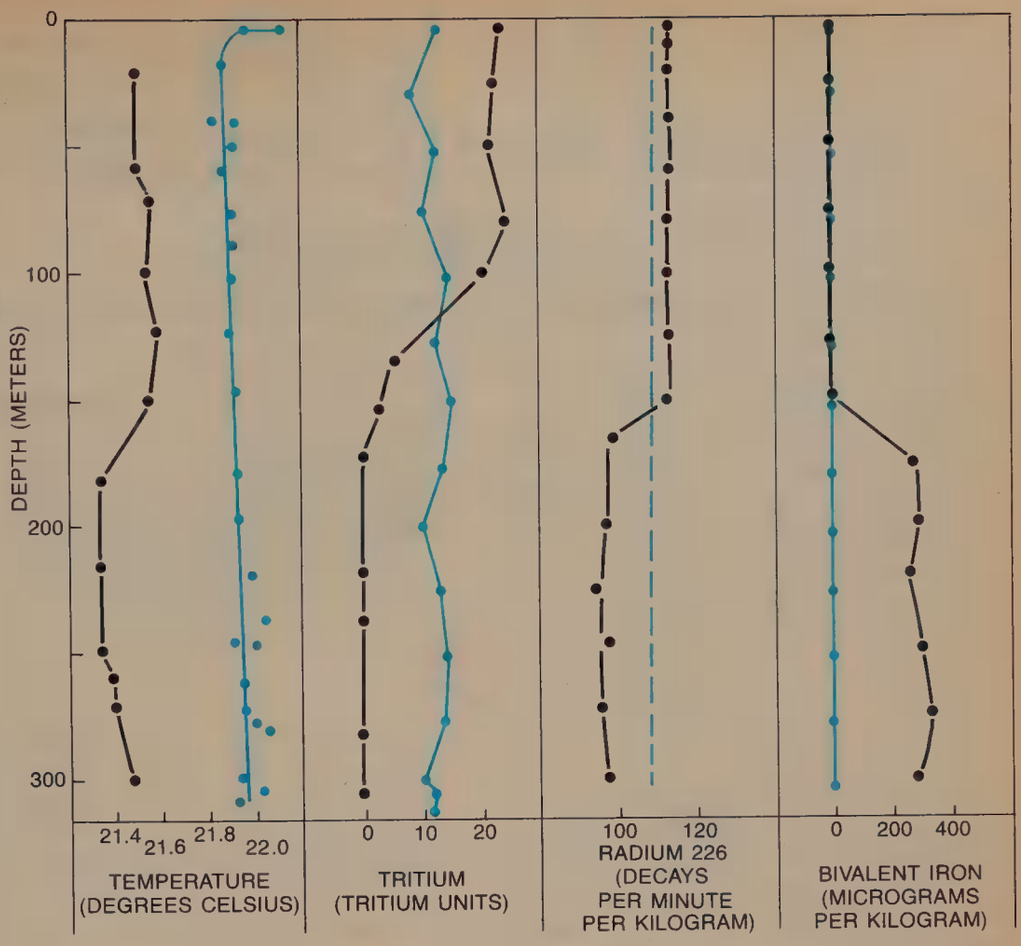
years ago, toward the end of the "little ice age" in the 18th century.

The calculations that are described above were based on the assumption that the deep fossil water constituted an absolutely closed system. In reality no such system can be completely closed; there is always some exchange by diffusion even across a pycnocline, or density boundary. To what extent, then, was the deep water truly isolated during the long-term stratification?

In 1967 Abraham Lerman of the Weizmann Institute calculated the rate of mixing across the pycnocline that could be expected simply on the basis of the random Brownian motion of molecules. He concluded that mixing by molecular diffusion would be a very slow process and would take thousands of years. Actually the picture is not quite so simple. There is seepage and there are springs at the bottom of the lake. On occasion a plume of dense surface brine from the southern basin may plunge deep into the northern basin. Temperature profiles in the deep-water mass and transition zone allow one to estimate the amount of influx of hot (and perhaps saline) water into the deep-water mass. In effect, then, the system is not truly sealed off. By measuring short-lived radioactive tracers such as tritium (which is introduced from the atmosphere) one can, however, calculate an upper limit to the rate of renewal of the fossil waters. It must be a slow process indeed, on a time scale of at least 1,000 years.

Even in the years immediately preceding the overturn (when the stability of stratification had decreased, increasing the chance that parcels of surface water would invade the deep-water mass), temperature and tritium measurements indicate that less than 2 percent of the fossil water was being renewed per year. It was only the final overturn in the winter of 1978–79 that brought water down from the surface to mix with the fossil water body.

What combination of conditions can bring about an overturn in Dead Sea waters? The essential requirement is undoubtedly a decrease in the water column's stability due to an increase in surface-water salinity resulting from a decline in the water level. For all this to happen conditions throughout the younger, upper layers have to be such that vertical mixing can take place. Such conditions evidently do not exist in the Dead Sea in the summer, when the surface layers are warmed, establishing a thermally controlled pycnocline. Moreover, in the winter or spring an influx of fresh water may dilute the surface brine and establish a pycnocline based on a salinity gradient. It is generally in the fall or early winter that a pycnocline near the surface is deepened by cooling of the surface layers;



MIXING OF FOSSIL WATER with surface water during the overturn is documented by vertical profiles of temperature and of the concentration of three chemical components. Measurements were made in January–February, 1978, one year before the overturn (*black*), and the next February, during the overturn (*color*). Broken lines indicate estimated data. The preoverturn profiles show a significant difference between the younger waters and the deep fossil water; such differences are no longer evident after the overturn. The long isolation of the fossil water body gave it characteristic chemical and radiochemical properties: the low values for radioactive tritium and radium and the presence of bivalent iron, which indicates a lack of oxygen.

the depth to which mixing takes place depends on the current state of the water column. Often one then encounters a hierarchy of layers of increasing age, ranging from very young water at the top to the fossil layer at the bottom.

It is only after a succession of dry seasons that the surface layers become salty enough (and thus dense enough) for the autumn mixing to reach all the way to the transition layer delineating the fossil water body and on to the bottom of the lake, and so bring about an overturn. A particularly rainy season may, on the other hand, cause a pycnocline to persist at a shallow depth for several years. After the wet season of 1979–80, for example, the water column (which had been homogenized during the overturn) was stratified again for three years, until the end of 1982. The three-year period was not, however, long enough to reestablish the "fossil" nature of the deep waters.

Have there been overturns in the past? In searching for a record of such events one might start with the fact that a period of relatively low water (and consequent high average salinity) is a necessary condition for overturn (although perhaps not a sufficient condition, since overturn can still be avoided by occasional flooding). During the recent overturn high salinity resulted in the precipitation of halite, the mineral form of sodium chloride, from the Dead Sea brine. Any previous overturn may have been accompanied by a similar precipitation event, which should be recorded by the presence of halite crystals in the Dead Sea's sediments. A careful search for layers of precipitated halite in the sedimentary column might yield evidence for dry periods that may have been associated with previous overturns in the long history of the Dead Sea.

As one might expect, the changes in salinity we have been recounting had a major impact on the Dead Sea's biota. In 1979, after years of low water and increasing salt concentration had led to the overturn, the lake seemed to be sterilized. Then came the rainy winter of 1979–80. The number of microorganisms increased dramatically, with up to 19 million cells per milliliter at the surface. The bloom of microorganisms changed the optical properties of the surface water. The lake changed from blue green to reddish blue, giving visual evidence that the Dead Sea is not quite dead.

The world's densest compute is now the heart of the world's mos 32-bit computer.

Our 32-bit CPU.

From time to time, miracles of technology come along to make previously impossible tasks not only possible, but easy. That tiny 450,000-transistor integrated circuit is one of those technological miracles.

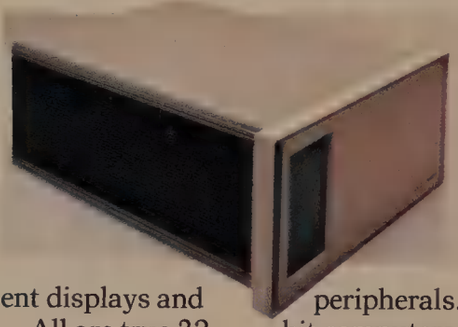
Hewlett-Packard didn't develop it just to break the record for most transistors on a chip, but to put on an engineer's or scientist's desk a computer so powerful that it can do the work of mainframes costing four times as much.

32-bit computers for 32-bit applications.

The new HP 9000 computer based on this and four other 'superchips' can handle formidable engineering and scientific problems. The scientist solving complex systems of equations, the mechanical engineer doing finite element analysis or three-dimensional modeling, the electrical engineer analyzing complex circuits or designing very large-scale integrated circuits — these are the kinds of technical people and problems the HP 9000 family is designed for.

It comes in three versions. The integrated workstation is complete with keyboard, color or monochromatic graphics display, fixed and flexible disc drives, and printer. For systems manufacturers, there's a rack-mountable box. And for a variety of single-user and multi-user applications, the minicabinet version works with many different displays and peripherals.

A rack-mountable version is available, too.



All are true 32-bit computers, with 32-bit CPUs, memories, and data paths. And the multi-CPU architecture lets you nearly double or triple your processing power at any time by adding one or two CPU boards. Without increasing the computer's size.

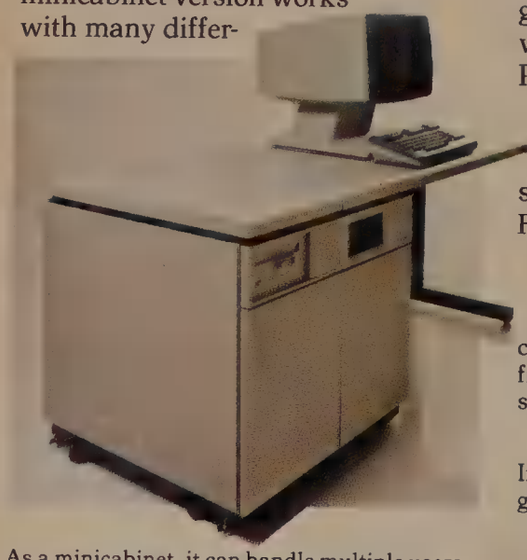
Two operating systems are better than one.

The integrated workstation is available with a choice of operating systems. One is HP's highly evolved, high-performance Enhanced BASIC, augmented with 3-D graphics and a software innovation called a run-time compiler. This substantially increases program execution speed, while retaining an interactive development environment.

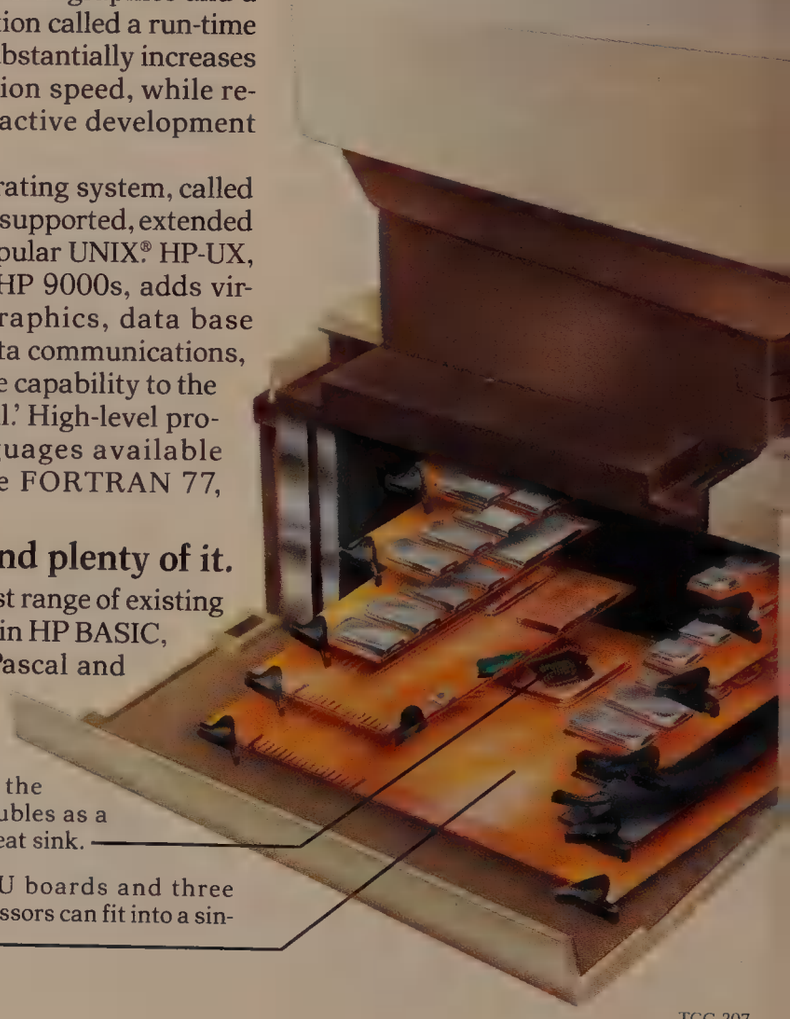
The other operating system, called HP-UX, is a fully supported, extended version of the popular UNIX® HP-UX, available on all HP 9000s, adds virtual memory, graphics, data base management, data communications, and enhanced file capability to the basic UNIX 'shell.' High-level programming languages available with HP-UX are FORTRAN 77, Pascal and C.

Software, and plenty of it.

Much of the vast range of existing software written in HP BASIC, FORTRAN 77, Pascal and C is transportable to the HP 9000. HP will also be offering proprietary software packages emphasizing computer-aided design and engineering. These will tie the HP 9000 into HP's Manufacturer's Productivity Network (MPN). Third-party software suppliers will be providing many of the most widely used CAE packages for 32-bit computer systems. And both HP 9000 operating



As a minicabinet, it can handle multiple users.



The 32-bit CPU chip is bonded to the finstrate which doubles as a signal carrier and heat sink.

Up to three CPU boards and three Input/Output Processors can fit into a single HP 9000.

TCG-207

hip
ffordable

systems offer extensive program development tools.

You also get a choice of communication tools. The HP 9000 is currently compatible with Ethernet,™ and with HP's Shared Resource Manager (SRM) which lets clusters of HP 9000 and 16-bit desktop computers share data and use common peripherals.

Links to central computers are also available. And in late 1983, HP will offer local area networks based on the IEEE-802 standard.

New technology from the silicon up.

The five superchips that make the HP 9000 possible are the 32-bit CPU, which can execute a million instructions per second; an eight-channel Input/Output processor (IOP); a random-access memory chip capable of storing 128K bits of data; a memory controller that 'heals' up to 32 bad memory locations; and an 18-megahertz clock.

Hewlett-Packard's advanced NMOS-III process makes it possible to put 450,000 transistors on a chip only 0.4 centimeters square. This tremendous density of electronic components could have required an expensive and elaborate cooling system.

Instead, HP engineers developed a new mounting structure called a finstrate, a copper-cored circuit board, which acts as both cooling fin and substrate. The finstrates containing the CPU, IOP, memory, and clock chips are housed in a lunchpail-sized module.

One user, one mainframe.

Clearly the trend in engineering and scientific computation is away from large machines shared by multiple users and towards networks of powerful personal workstations, sharing peripherals and data bases. The reason is compelling. An engineer or scientist in personal control of an HP 9000 can solve so many more problems more easily that the increased productivity alone makes the cost of individual computers easy to justify.

For complete information about this powerful breakthrough in 32-bit computing, contact the local HP sales office listed in your telephone directory. Ask a Technical Computer Specialist for a demonstration. Or write to Pete Hamilton, Dept. 41151, Hewlett-Packard, 3404 East Harmony Road, Fort Collins, CO 80525.



HEWLETT PACKARD

UNIX is a trademark of Bell Laboratories. Ethernet is a trademark of Xerox Corporation.



This illustration of an optical fibre cable was created for Northern Telecom's Optical Systems Division by Acorn Technical Art Studio Inc., Toronto. Northern Telecom's unique star-shaped optical fibre cable is popular with telephone companies and specialized common carriers and is in operation throughout major telecommunications networks in Canada, the United States and the Caribbean.

CANADIAN TELECOMMUNICATIONS

Nearly 80% of Canada's 25 million people are spread along a narrow but very long corridor (6,400 km) just north of the U.S. border, with the remaining 20% thinly spread all the way to the Arctic Ocean. According to R.D. Sloane, President of TransCanada Telephone System (TCTS): "The highly sophisticated [communications] network that serves Canadians today was historically a response to the stimulus and imperatives of geography. Canadians have always felt the need to talk frequently to one another, to overcome distance and to avoid the consequences of personal and regional isolation."

In what Sloane calls "a unique blend of monopoly and competition," ten major telephone companies provide 94% of Canada's telephone service and together form the umbrella organization, TCTS. Six of these companies are investor-owned, three are operated by provincial governments, and the tenth, Telesat Canada, which provides domestic satellite services, is a federal Crown Corporation.

It is thus not surprising that there is federal participation in most of the major technological development being mounted by the Canadian telecommunications industry.

There are currently three major thrusts of development in Canadian communications technology:

(1) The creation of an all-digital communications network across Canada adapted to the simultaneous transmission of voice, data, and other communications services (an "Integrated Services Data Network," or ISDN).

(2) Continuing development of satellite communications, following Canada's establishment in 1972 of the world's first domestic satellite communications system.

(3) Pioneering of optical fibre transmission media for land-line voice and data transmission.

The move to integrated digital networks

The Canadian telecommunications industry, as stated by Sloane, is divided into areas of monopoly and open competition. Its two major intra-Canadian operating components are TCTS and CNCP Telecommunications. TCTS is responsible for the nation's

Advertisement

We built our reputation from the ground up







Designing and operating a reliable satellite system for one of the world's largest countries presents many unique challenges. Telesat Canada has met every challenge with innovation and has been involved in the design of a variety of systems:

- FDMA Message
- SCPC
- TDMA
- TV and Radio program services
- DBS

Our reputation for technical excellence coupled with operating experience in these services has spread beyond the borders of Canada. Today, international consulting is a growing and important part of our business. Consulting services are offered in system design and program management in the following areas:

- satellite systems
- earth stations
- communications networks
- satellite control systems
- operations and training

In space or on the ground, our commitment to excellence and innovation has earned us international recognition as the company on the leading edge of satellite communications technology. For more information on Telesat Canada and our consulting capabilities, contact:

Consulting Services
Telesat Canada
333 River Road, Ottawa
Ontario, Canada K1L 8B9
Tel: (613) 746-5920 ext. 359

TeleSat

Telesat Canada

New Horizons in Communications



long distance telephone network, and its members provide most of the public local telephone service in Canada. CNCP Telecommunications, which began as a consortium of the two major railroads in Canada, is responsible for the provision of public telegram service. Teleglobe Canada, a federal Crown Corporation, provides international telecommunication services, except for Canada-USA traffic, which flows over an integrated North American network.

Beyond these basic monopolies, strong competition exists for the carriage of radio and television signals and data and computer communications.

The potential for enhanced speed and reliability of digital transmission has long intrigued Canada's communications planners. Digital data transmission for inter-office applications was made available in 1965, via pulse code modulation (PCM) of analog signals. By 1975, North America's longest (700 km) high capacity broadband digital transmission system using 12 coaxial copper tubes, was in place between Montreal, Ottawa, and Toronto. This was the precursor of "The Dataroute," a transcontinental broadband data highway with links to the USA at major border crossing points. This year, an advanced digital radio system was completed across the country (6400 km) by TCTS, using ten working and two standby radio channels providing 1344 voice-equivalent channels operating at 90 megabits per second.

The availability of digital transmission has made possible the goal of integrating voice, data, and other communications services. Says Robert Haughton, Director International of TCTS, "Integration has two components: one applicable to the fusion of digital transmission and digital switching into a single . . . system dedicated to a particular service such as telephone or data – referred to as the Integrated Digital Network (IDN) – and the other focussing on the provision of different services over a single digital channel – the Integrated Services Digital Network (ISDN).

"Progress towards all-digital long distance networks is expected to be relatively rapid: within the next decade about one half the nation's long distance circuits should be digital . . . and by the end of the century perhaps 90% or better."

The ISDN concept, in which a single channel can be used to carry various services (such as video, telephony, and data transmission) over the same loop, is now being intensively investigated by the Canadian industry. One key to the requirement for broadening bandwidth and decreasing signal attenuation, inherent in ISDN service, is the introduction of fibre optic cable, first in transmission trunks, and then into the customer loop.

Fibre optics: key to the communications of tomorrow

Fibre optic systems are defined by Terence Heenan of AEL Microtel Ltd. as "a combination of hair-thin glass fibres and electronics with the capability of carrying over 4,000 voice circuits per glass fibre pair ... the state of the art method for transmitting sound, video, and data simultaneously by using high speed streams of laser generated light pulses over glass fibres."

Initial experiments in the use of optical fibres, carrying signals in the form of light pulses through transparent fibres, were carried out by Bell Telephone starting in 1976. The first major field trial is being carried out in a rural area, surrounding the village of Elie, Manitoba.

Meyer Akgun of the federal Department of Communications makes the point that: "People living in rural areas have actually a greater need for communications to satisfy their information needs compared to urban dwellers who have easier access to such services. In recognition of this situation, the . . . Government of Canada initiated a rural communications study program. Among many approaches . . . the application of fibre optics technology to an integrated services subscriber access network proved to be one of the most promising."

While the trial was completed this March after 18 months of operation, its performance was so satisfactory that it was decided to continue the operation as a test-bed for future technology and service trials.

In 1980, the Saskatchewan Telephone Co. began to install a 3,400 km fibre optics broad band network linking the province's major population centres. Now completed, it is the world's longest commercial fibre optics network, providing full voice and data services, as well as cable television signals to 52 communities.

At Regina, the provincial capital, Northern Telecom has built a $14 million fibre optics manufacturing facility occupying 91,000 square feet. The company sold $37 million worth of fibre optic transmission systems in 1982, and this year announced its first contract for single-mode optic fibres. Single-mode fibre offers three to five times the number of communications channels as multi-mode fibres.

Satellites: bringing people and continents together

Canada, with her problems of a population thinly scattered over a vast area, saw early the advantages of satellite relaying of communications. Telesat Canada, the federal Crown Corporation, was founded in 1969. In 1972, the launch of the Anik A satellite established the world's first geostationary space satellite in domestic communications service.

R.M. Lester of Telesat explains the inherent simplicity and effectiveness of the satellite system for distribution of broadcast signals this way:

"To reach an unlimited number of television receivers scattered randomly throughout a service area, requires the use of only one satellite earth station transmitting to a satellite in-orbit. This in turn transmits the signal back to earth where it can be received over an area which may be regional or country-wide in scope. Satellite television transmissions have utilized FM techniques where the broadband satellite channels have had sufficient power to support excellent picture quality while requiring only modestly sized receive antennas ranging from 1.2-4.5 metres in diameter."

The Telesat reliability factor is better than 99%, providing such service as global TV relay with local-area distribution, telephone service, computer data transfer, teletype, and facsimile. Canada now has over 100 earth stations in operation and more communications satellites than any other country except the USA and the USSR.

The development of Canadian satellite technology has given several Canadian firms an entry into a significant and growing world market. Spar Aerospace Ltd., for instance, employs 2,000 people, and did $178 million worth of business last year in the aerospace, communications, and military markets. The company received worldwide attention for its provision of the "Canadarm," the manipulator system for the US space shuttle program, which will play an important role in future in the orbiting, retrieval, and repair of space satellites. Last year, the company signed its largest contract yet for $125 million with the Brazilian state company Embratel for two satellites of the Anik "D" type, which can each carry 16,000 two-way voice channels or 24 TV channels. This will be the first domestic communications satellite system in Latin America.

AEL Microtel Ltd., a subsidiary of British Columbia Telephone Co., has just sold its "Spacetel" system to Manitoba Telephone. The central control system in Winnipeg will connect remote Indian settlements in the sub-arctic north of the province to the North American network, via light-weight terminals. As described by Microtel president Terence Heenan, Spacetel is: "A thin-route

Advertisement



NEW HORIZONS: ONTARIO/CANADA

The growth of the high technology industry is happening faster in Ontario than just about anywhere in the world today.

For example, the research, development and manufacturing of the Spar Remote Manipulator "Space Arm" and the Telidon Videotex System, are just two of many major developments taking place in Ontario.

Another exciting project is the establishment of the Microelectronics Technology Centre in Ottawa which will provide a design, production and testing capability for our high tech industry.

We think the technological innovation now going on in Canada has a lot to do with the kind of tomorrow all of us will be experiencing. And the more our high tech community grows, the greater our collective contribution to that tomorrow will be.

If you, or your company would like to be part of the new horizons Ontario is developing today, contact us.

ONTARIO Ontario **CANADA**

Honourable Frank S. Miller
Minister of Industry and Trade

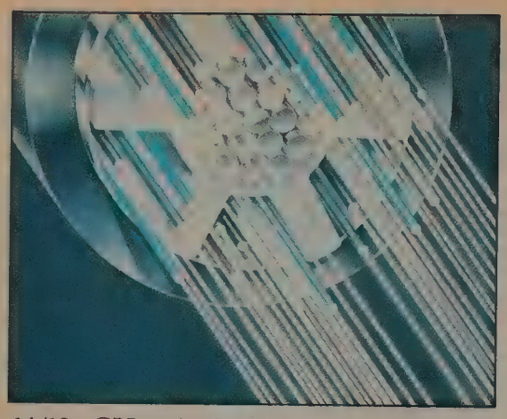
Honourable William G. Davis
Premier of Ontario

Ontario House
Rue Royale 97-99
Brussels,
Belgium
Telephone: 32-2-218-0600
Telex: 46-63559

Ontario House
Charles II Street
London SW1Y 4QS
England
Telephone: 44-1-930-6404
Telex: 51-262517 ONTLDN

Ontario House
109, rue du faubourg St.-Honoré
Paris
France
Telephone: 33-1-563-1634
Telex: 42-650865F ONTPAR

Government of Ontario, Canada
Bockenheimer Landstr. 51/53
D-6000 Frankfurt/Main
West Germany
Telephone: 49-611-72-0971
Telex: 41-416176 ONFRA



14/12 GHz single channel per carrier (SCPC) system, configured as a star network to extend subscriber services available at a central site to any location within the appropriate satellite footprint. High quality voice transmission is provided using 32 kilobit per second digital encoding and quaternary phase shift keyed modulation."

In application, subscribers in the isolated area are assigned a metropolitan exchange number, receive dial tone from that centre, and gain access to all telephone services available, as well as broadcast services.

Telephony for the third world

The emphasis on improvement of rural services, and for adapting high-technology equipment to low-technology environments, is a major basis for the increasing trade in communications hardware and software between Canada and the Third World.

Bell Canada's assistant vice president (Engineering), R.C. Terrault, describes his company's recently completed Non-Urban Service Improvement program (NUSI) this way:

"Prior to the program, many of Bell's rural subscribers shared their telephone lines with up to seven or eight other customers; 450,000, or nine per cent of all main telephones, were located outside urban areas, and of those, 250,000 were in multi-party service . . . The program meant that urban areas were extended to provide urban service to expanding concentrations of customers, and locality rate areas for islands of denser population were established wherever needed; the introduction of DMS-1 digital multiplexers and rearrangement of telephone facilities was required throughout Bell territory to provide all customers with access to four party lines, and, in many cases, with individual or two-party lines. At the conclusion of the program in 1981 the number of multi-party line services totalled approximately 200,000 and the number of subscribers per line was reduced to an average of 2.5 from 4.6."

Such a program is evidence of the strong sense of social responsibility which has been a tradition of the Canadian telephone industry. (It may be remembered that Alexander Graham Bell's invention of the telephone originated from his work in trying to alleviate the problems of deafness.) J.S. McNeice of Bell Canada gives this short list of recent "action taken" on telecommunications problems of the handicapped:

"* Centres for Special Needs were established in two major Canadian cities with staff specially trained in the communications needs of the disabled.

* A standardized portable keyboard device was developed for use by the deaf, with a special Assistance Centre for dealing with calls from keyboard users.

* A small voice synthesizer was produced to convert visual display material to audio output, in order to facilitate the employment of blind operators.

* A "handsfree" telephone was developed to dial the operator through the use of switches designed for paraplegics."

Global business pushes toward office automation

The Canadian communications industry, of course, did not grow and thrive on social responsibility alone. The message for business is well expressed by A.J. de Grandpre, Chairman, President and Chief Executive Officer of Bell Canada Enterprises, Inc.:

"It is the senior executive in business organizations today on whom the microelectronic era may well have the greatest impact – not the secretary, the clerk or the factory worker. While these workers have to learn new techniques, new Information Age technology is forcing on executives an entirely new way of thinking and new ways of doing business."

With the increasing density of business information available, Oscar Stubits of CNCP Telecommunications reports, "there is an increasing demand for new services. This demand is, to a large degree, based on the need to provide the already large installed and rapidly increasing base of 'dumb' and intelligent terminals with the capacity to communicate between them. Most recently, we introduced Teletax, electronic mail, and an asynchronous communications service for word processors."

The demand for new services means a correlative demand for new products. Mitel Corp. is a ten year old company that has made an outstanding success in developing PBX exchanges for business, to which it is now adding work stations and advanced business telephones called "Supersets." The company has grown from two to 5,000 employees, with $250 million in annual sales and manufacturing plants in the USA, Mexico, Ireland, the United Kingdom, and West Germany.

One of Mitel's design emphases is to configure its systems for compatibility with the major existing data processing systems.

Mitel's vice-president, marketing, James Mackie, warns, however, of the dangers of misapplication of office technology. He quotes the aphorism "Yes, I know office technology is the answer – but what is the question?" Frequently, he believes, management makes the right decisions on office automation for the wrong reasons. Thus, video conferencing will bring great benefits to far-flung corporations – but it won't obviate the need for personal travel, any more than the telephone has eliminated the importance of personal contact in business.

Cellular radio systems being developed

One new development of communications technology which has immense potential for both business and private use is the development of "cellular" systems for mobile telephony. While telephone companies in some countries see this as a possible infringement of their communications jurisdiction (as earlier Citizens' Band communications were also sometimes regarded), Canadian telephone systems are in the forefront of developing "cellular" radio links.

NovaTel's Dr. Aremus Slekys describes the "Aurora," an "AUtomatic ROaming RAdio" system developed by NovaTel:

"Aurora is a medium to high capacity cellular mobile telephone system . . . developed to satisfy the widely varying needs of both public and private users in the Province of Alberta – one of the largest users of mobiles in North America, with 30,000 units in service in 1983.

"The system is designed to interface directly with the nearest telephone exchange thus eliminating the need for an overlay network, dedicated central switches and extensive trunking. Computer control of the overall system forms a heirarchy beginning with the Master Mobile Centre followed by Regional and Local Mobile Centres. Aurora has the flexibility to be compatible with numerous signalling protocols, national or international standards, and can be implemented economically over a wide range of frequency bands."

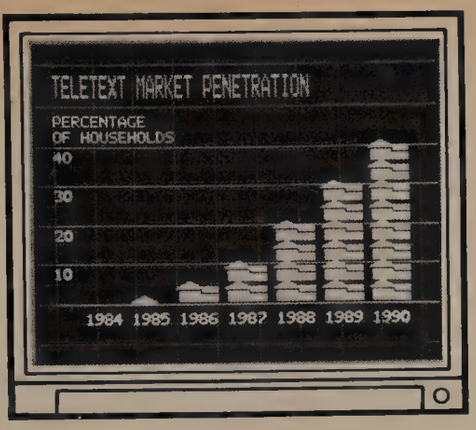
Playing a role in a worldwide revolution

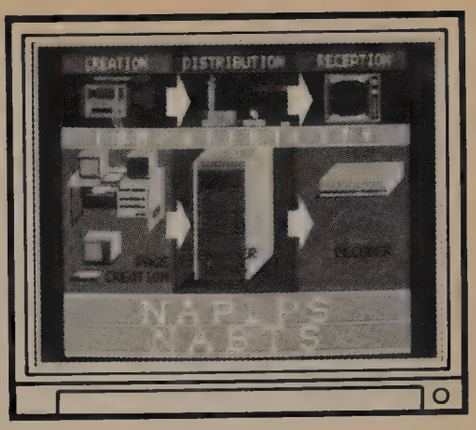
Telecom 83, a world showcase on telecommunications technology, will be taking place at Geneva, Switzerland from Oct. 26 to Nov. 1. The meeting represents one of the key focal points to "World Communications Year" as designated by the United Nations.

Telecom 83 is the fourth in a series of quadrennial World Telecommunications Exhibitions organized by the International Telecommunication Union (ITU), a UN agency representing 146 countries. The meeting is expected to be the largest so far, with 650 exhibitors from 40 countries. The exhibition is accompanied by a prestigious World Telecommunications Forum, with

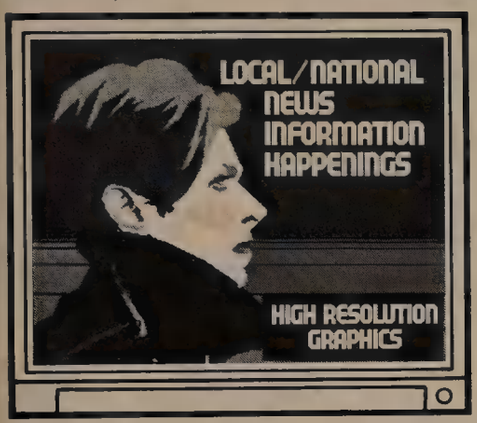
Advertisement







Telidon, the world's most advanced videotex computer protocol was developed in Canada.

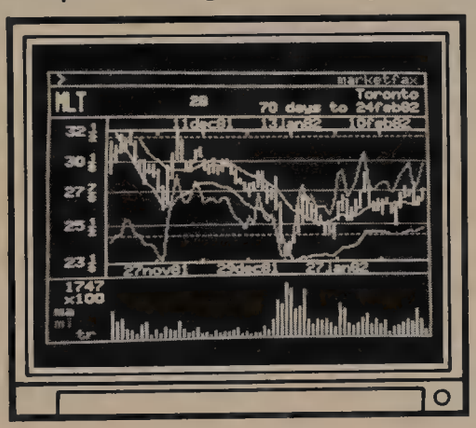






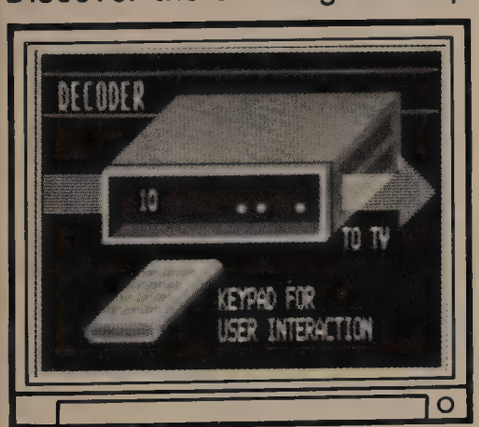
Today, Canadian companies continue pioneering technological advances in electronic information.







Discover the exciting developments in information technology and telecommunications







at Telecom 83, Hall 2, Stands 117-145, Palais des Expositions et des Congrès de Genève.

Wherever you live the Canadian embassy, consulate or High Commission near you can provide information about Canadian companies and their capabilities in telecommunications.

Canada

External Affairs Canada
Affaires extérieures Canada



nearly 200 carefully selected original papers presented in sessions devoted to the technology, economics, and legal aspects of telecommunications.

While the USA boasts the largest single national pavilion, with 80 exhibitors crowding 64,000 square feet of floor space, its northern neighbour, Canada, is taking an impressively major part in both the exhibits and technical proceedings.

Canada's Federal Department of External Affairs is coordinating a national pavilion of 15,000 square feet, in which 17 manufacturers are participating. Four large companies, Bell Canada, Northern Telecom, Teleglobe Canada, and Mitel, also have their own exhibit space, aggregating almost as much area again.

In the technical Forum, Canada's contribution is also disproportionate to its size in terms of population. Canadian communications organizations submitted 20 abstracts of proposed papers, of which seven were accepted for presentation, and another six for publication in the printed Proceedings of the Forum. R.D. Sloane, President of Trans-Canada Telephone Systems and A.J. de Grandpre, Chairman, President, C.E.O. of Bell Enterprises Inc., were chosen as world telecommunications leaders to make official opening presentations. R.N.E. Haughton, Director, International of TCTS, was chosen to chair the Mobile Communications session.

The Papers Committee reported that Canada is presenting eight per cent of the total Forum papers, and has the highest ratio of submissions accepted for presentation: 35%, vs 29% for Germany and Japan and 21.4% for the USA. (All papers are presented in simultaneous translation in English, French, and Spanish.) This showing reflects the importance of the telecommunications industry to Canada.

Summing up the raison d'etre of Canada's participation in Telecom 83, Douglas Sloane of TCTS says:

"Rapid advances in (telecommunications) technology are now generating a worldwide revolution in which Canada has a chance to play a leading role. This is an opportunity which, properly exploited, can create new jobs, improve trade balances, and make a significant contribution to the economic recovery of this country."

Canadian Papers Accepted for Presentation at Telecom 83

A Fibre Optic Integrated Services Field Trial in Canada: M.B. Akgun (DOC), K.B. Harris (Cdn. Telecommunications Carriers Assoc. – CTCA); G.A. Tough (Manitoba Telephone System). Rural Networks Session.

Planning for the Introduction of New Digital Customer Services and Loop Technologies: F.M. Banks (Bell Canada); G. Gara (Bell Northern Research). ISDN Session.

An Integrated System Implementation in Canada: O. Stubits (CNCP Telecommunications). Enhanced Services Session.

Technology and Social Responsibility: J.S. McNeice (Bell Canada). Terminal Equipment Session.

Aurora – An Automatic Roaming Radio System: A. Slekys (Alberta Gov't Telephones). Mobile Systems Session.

Software Aids for National Architecture Planning: D. MacDonald; M.O'Leary (Bell-Northern Research). Technology Trends Session.

Satellite Applications in Canadian Broadcasting: R.M. Lester (Telesat Canada). CATV-Broadcasting Session.

For Publication Only

Network-based Data Services of TCTS: A.M. Rybczynski (TCTS).

Integration of Satellite and Terrestrial Digital Technologies for Business: P. Kiar (Bell Canada).

A Single Channel Per Carrier Satellite System: S. Hussain (AEL Microtel).

Optimizing Rural Telephone Networks: R.C. Terrault (Bell Canada International).

Microwave Frequency Coordination in Canada: M. Derzai (CTCA).

Optimizing Management Training and Development in a Changing Technical Environment: E.J. Clifford (Bell Canada).

Canadian Companies Exhibiting at Telecom 83

AEI Telecommunications Canada Ltd., Winnipeg, Manitoba. *(Small and large telephone switching systems.)*

AEL Microtel Ltd., Burnaby, British Columbia. *(Satellite and other receiving and multiplexing systems, advanced R&D).*

Allcom Data Ltd., Nepean, Ontario. *(Data Communications Products.)*

Bell Canada International, Ottawa, Ontario. *(Major telecommunications system consultants and designers).*

Canadian General Electric Ltd., Toronto, Ontario. *(Fibreglass and other telephone conduit systems.)*

Canadian Marconi Co., Montreal, Quebec. *(Data switching and radio systems.)*

Century III Electronics Inc., Vancouver, British Columbia. *(Broadband communications networks.)*

Croven Crystals, Whitby, Ontario. *(High frequency crystals for communications and radar systems.)*

Dept. of Communications, Ottawa, Ontario. *(Federal government agency responsible for all media of communications in Canada.)*

Farinon Canada Ltd., Dorval, Quebec. *(Analog and digital microwave radios.)*

Gandalf Data Ltd., Ottawa, Ontario. *(Data communications products.)*

Global Thermoelectric Power Systems Ltd., Bassano, Alberta. *(Thermoelectric power products.)*

Hermes Electronics Ltd., Dartmouth, Nova Scotia. *(Antennas and other communications equipment.)*

Idacom Electronics Ltd., Edmonton, Alberta. *(Data communications test equipment.)*

Memotec Data Inc., Montreal, Quebec. *(X.25 communication processors and multiplexers.)*

Mitel Corp., Kanata, Ontario. *(PABX switching systems, work stations, telephone handsets.)*

Northern Telecom Ltd., Mississauga, Ontario. *(Manufacturers of complete line of telephone utility and subscriber equipment.)*

Novatel Communications Ltd., Montreal, Quebec. *(Cellular mobile telephone systems.)*

Spar Aerospace Ltd., Ste. Anne de Bellevue, Quebec. *(Satellite and aerospace systems.)*

Spillsbury Communications Ltd., Vancouver, British Columbia. *(Specialized radio equipment.)*

SR Telecom Inc., St. Laurent, Quebec. *(Urban and rural subscriber radio systems.)*

Teleglobe Canada, Montreal, Quebec. *(Crown Corporation responsible for overseas communications services.)*

Hugh C. McIntyre, the author of this report, is a technical writer based in Toronto who has covered developments in computers and communications over the last 20 years.

Art direction, supervision of text and illustration and coordination of production were provided by Camp Associates Advertising Limited, Toronto.

Advertisement



Around 5,000 years ago, Sumerian messages were written on clay tablets. Some even came with their own clay envelope.

"...the globe has contracted spatially, into a single large village."

Marshall McLuhan
The Gutenberg Galaxy

In ancient times, a message could travel only as fast as the messenger.

Today, thoughts and words are flashed around the globe electronically, at the speed of light.

Teleglobe Canada carries oversea messages to and from Canada, allowing Canadians to trade information with people in more than 80 countries as easily as with their next door neighbours.

Behind this ease of communication is a vast system of satellites and undersea cables linking the continents. Teleglobe Canada operates the Canadian part of this sophisticated system, providing telecommunications to the four corners of the earth.

Teleglobe Canada

Bringing people and continents together

Six Millenniums of Buffalo Kills

At a site in western Canada named Head-Smashed-In, Indian hunters slaughtered buffalo by stampeding the herd over a cliff. They started in 3700 B.C. and continued until well after the white men had come

by B. O. K. Reeves

According to many early 19th-century accounts and illustrations, the Indians of the Great Plains of North America hunted buffalo by riding them down on horseback. The implication is that these peoples were not successful buffalo hunters until some time after Europeans introduced horses into North America in the 16th century. Archaeological work in the wide grasslands east of the Rocky Mountains contradicts this assumption. From the time the first Paleo-Indians entered the region some 10,000 years ago, they were redoubtable hunters even though they hunted on foot. Mammoths were the game they favored at the start, and when those great beasts became scarce, they turned to hunting two different species of buffalo: the large, straight-horned *Bison antiquus* and the smaller *Bison occidentalis,* both now extinct. The Paleo-Indians' successors on the plains continued to hunt buffalo on foot, and as a communal enterprise.

Here I shall describe what has been unearthed at one such communal bison-hunting site, 130 kilometers south of Calgary at the southeastern edge of the Porcupine Hills in western Alberta. There the Piegan Indians of historical times and their predecessors have been stampeding buffalo to their death over a cliff for at least 5,600 years and possibly as long as 9,000. As with other "buffalo jump" sites in Canada and the U.S., the topography of the site favored group stalking and stampeding of buffalo herds. In their language the Piegan call the place Head-Smashed-In.

The cliff the hunters used as a jump is a sandstone outcrop that had weathered and collapsed over the centuries so that a talus slope of detached sandstone slabs and smaller fragments built up along its foot. Today the vertical drop from the cliff edge to the top of the slope measures between nine and 10 meters. Westward, back from the edge of the cliff, the valleys of six tributaries to a central stream, Olsen Creek, occupy a shallow basin, some 36 square kilometers in extent, rimmed by a continuous highland area except where Olsen Creek leaves the basin. More than 500 cairns, rock piles about a meter in diameter and a third of a meter high, still mark paths as much as eight kilometers in length along which the communal hunters drove the buffalo to the cliff.

The talus slope below the cliff was not formed by erosion alone. Most of its bulk is the result of the detachment from the cliff face and subsequent downhill movement of two separate massive blocks of sandstone. The first of these block falls, called slumps by geologists, came in about 3700 B.C. and the second at about the start of the Christian Era. It is difficult to estimate how high the cliff stood above the talus after the first slump, but even after the second it was nearly twice as high as it is today.

Stampeding buffalo over the edge of a cliff was not the only hunting method pursued by the Plains Indians of the fourth millennium B.C. To judge by the archaeological evidence and by Piegan practices in historical times, the buffalo were sometimes driven into bogs, arroyos and box canyons and even into corrals or pounds built by the hunters for the purpose. The first white man to visit Piegan country, a Hudson's Bay Company trader named Peter Fidler, left a vivid account of pound drives he witnessed at Stimson Creek in December, 1797. "Our Chief is the Pound Master," Fidler wrote, "and whenever a herd [of buffalo] comes into the pound he must go & kill the first one. He generally shoots his one with a gun. The young men kill the rest with arrows, bayonets tied upon the end of a pole &c. The hatchet is frequently used & it is shocking to see the poor animals thus pent up without any way of escaping.... Some with a stroke of an axe will open nearly the whole side of a buffalo & the poor animal runs some times a considerable while all thro' the pound with all its internals dragging on the ground & trod out by the others."

In the course of a six-week stay at the pound, which measured 30 by 45 meters, Fidler estimated that the Piegan had trapped and butchered more than 250 buffalo. They could have taken more, he wrote, except that "when the wind happened to blow from the pound in the direction of the tents, there was an intollerable stench of the great number of petrified [*sic*] carcasses, &c on which account the reason of our leaving it."

Junius B. Bird of the American Museum of Natural History was the first professional to dig at Head-Smashed-In, but the results of his work in 1938 were never published. In 1949 Boyd Wettlaufer, then a graduate student at the University of New Mexico, ran test trenches both into the talus slope and at a campsite some 18 meters below the jump where the prime parts of the slaughtered buffalo had been carried for more detailed butchering. The Province of Alberta erected a cairn at Head-Smashed-In to commemorate his investigation, which was thought to be the first in Alberta. The cairn only attracted illicit excavators. Their appetite was further whetted by the news of an excavation in 1958 at a nearby jump site, "Old Women's," where the remnants of hunting activity had accumulated to a depth of more than seven meters and the earliest strata dated back to the first century A.D. By 1964 robbers had stripped the surface of Head-Smashed-In over an area of 1,400 square meters, digging pits as big as two meters in diameter and three meters deep.

With the integrity of the site in jeopardy the Glenbow-Alberta Institute and the University of Calgary joined forces to mount a major rescue effort over two seasons (1965–66). In the three months of the first season my associates and I worked mainly on the south part of the talus, opening up an area three by six meters in extent. At a depth of eight meters the area was reduced to a shaft 1.5 meters square, which was continued to 11 meters. Below the eight-meter level the shaft passed through a two-meter stratum of sterile fill before we encountered further traces of buffalo hunting. The reduced dimensions of the shaft

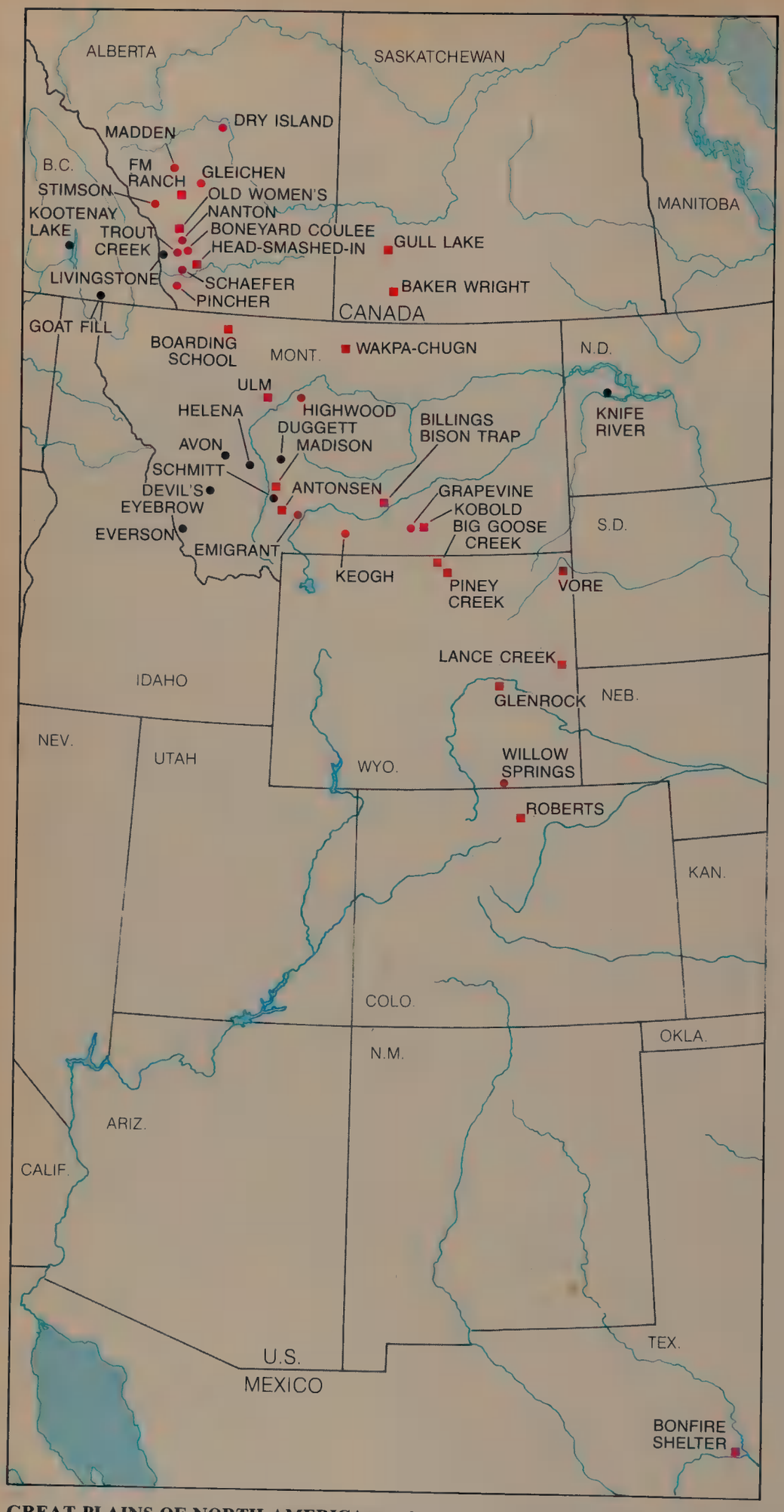
kept us from acquiring either bones or projectile points at this lowest level in quantities adequate for establishing phases of occupation and dates. The more plentiful material from the eight-meter level, however, dated back to about 1000 B.C., and we were confident that the bottom strata would prove to be considerably older.

In the second season our main effort was concentrated on the talus some 100 meters to the north of the 1965 shaft. One of the excavation units opened there, three by three meters in area, was excavated to a depth of four meters before slump-block sandstone was reached. The bottom stratum—a layer of sand, gray gold in color, from half a meter to one meter thick—held a bed of buffalo bones. Carbon-14 analysis of the protein collagen from the bones yielded an age determination of 3710 ± 100 B.C. Associated with the bones were side-notched projectile points that, although they were made from local stone, are typical of those first found some years ago at a Wyoming site known as Mummy Cave, 850 kilometers to the south.

The earliest slump block, which underlies the oldest bone strata in both the northern and the southern areas of the talus slope, consists of a 10-foot thickness of sandstone. Canadian government geologists, who helped us with our excavation in the 1966 season, trenched down to buried soils that had washed out from under the block. In the outwash they came across simple cobble tools and butchered buffalo bone. Both because of the depth of burial and be-



PRAIRIE VISTA occupies the middle distance in this view of the southeastern edge of the Porcupine Hills in Alberta. The diggers in the left foreground are members of the author's party who examined a Plains Indian buffalo-jump site here. The Indians who killed the buffalo made some of their stone points and knives out of pebbles and cobbles of chert and quartzite unearthed from the prairie's glacial till. Other tools, however, were made from exotic stones such as chalcedony brought from the Rocky Mountains, seen on the horizon.



GREAT PLAINS OF NORTH AMERICA are the scene of numerous buffalo-hunting sites discovered by archaeologists in recent years. They date from about 8000 B.C., in Paleo-Indian times, to post-Paleo-Indian periods that include the arrival of European settlers. Excavated and unexcavated buffalo-jump sites in Canada and the U.S. are respectively indicated by colored squares and dots. The sources of the kind of stone used to manufacture points and other tools by the occupants of the Alberta jump site, Head-Smashed-In, are indicated by black dots.

cause of the instability of the geologists' trench our group could not exploit their find. The geologists' deep trenching did not yield enough buffalo bone for carbon-14 analysis, and so the exact antiquity of these strata, evidently the lowest at Head-Smashed-In, remains undetermined.

One possible clue to their age comes from a chance discovery by Wettlaufer in 1948. Among the surface finds in his single season's work was a Paleo-Indian projectile point. It had evidently been an inclusion in soil dredged up during the construction of a water reservoir for the cattle that now graze the area. If the point, unmistakably of the Paleo-Indian tradition named the Cody Complex, was originally associated with outwash from under the earliest slump block, then Head-Smashed-In may have been a buffalo jump as long ago as the sixth millennium B.C.

In order of decreasing age the cultural debris we unearthed in our two seasons' work (and in an additional month's work at the site in 1972) provides a record of consistent although not continuous use of the buffalo jump from at least 3700 B.C. until the beginning of the 19th century A.D. Let me summarize the archaeological evidence for the use of the site by four successive prehistoric cultures and for its final use by the buffalo-hunting Piegan of modern times. To flesh out the bare bones of the archaeological record I shall comment on what the artifacts tell us about exchange networks and possible shifts of population over the same period, starting with the presence at the site of members of the Mummy Cave culture between 3700 and 3100 B.C.

Two preliminary notes, one climatic and one geologic, will be of help in interpreting what follows. First, studies of paleoclimate have established that the time between roughly 5500 and 3000 B.C., known as the Altithermal, was a period when conditions were generally drier and somewhat warmer than they were either before or after. Because the prairies are semiarid even today, it has long been thought that the earliest Americans would have shunned the plains in this period of even greater aridity. Such was apparently not the case. Quite apart from the early evidence at Head-Smashed-In, cultural debris has now been found at various Wyoming box-canyon buffalo-trap sites that date between 4700 and 4000 B.C. and belong to the Mummy Cave cultural tradition. Other evidence that the Altithermal was not a period of inactivity on the prairies comes from sites in Crowsnest Pass in the Rockies, 90 kilometers west of Head-Smashed-In. There a series of buffalo kills, accomplished by stampeding the herds into boggy ground, have been dated from 5200 to 2800 B.C.: a span of time that extends from the late Paleo-In-

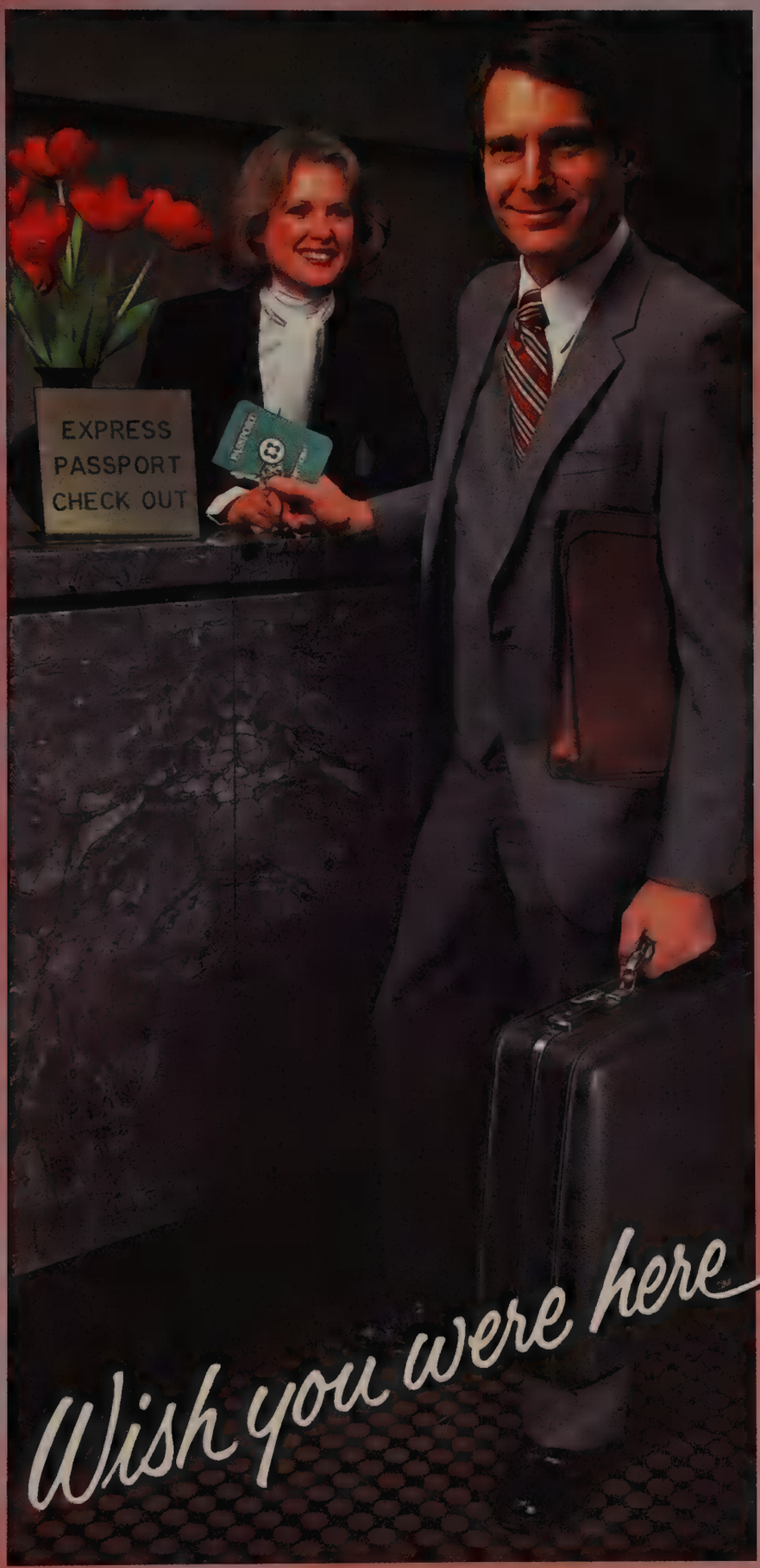
Why we make checking out of a Hyatt just as pleasant as staying at one.

The way we figure it, we give you so many extra touches at Hyatt, it would be a pity to have a long check-out line spoil them for you.

That's why when you check out of a Hyatt you simply hand in the passport you received when you checked in. You're checked out. Your bill is mailed to you within 24 hours, so there's absolutely no waiting in line. Absolutely no way you can forget how pleasant your stay was.

Any hotel can give you the essentials, but it's those extra touches ... like express check-out, that make the trip.

For reservations at 109 hotels worldwide, call your travel planner or 800 228 9000.



Atlanta
Discover magnificent Peachtree Center, 15 minutes from Atlanta International Airport.

Chicago
In Illinois Center overlooking Michigan Avenue and Lake Michigan.

Kansas City
At Crown Center, Hyatt has health club, tennis courts, pool.

Minneapolis
Hyatt is downtown on Nicollet Mall just 2 blocks from corporate business district.

Washington, D.C.
Walk to Capitol Hill, government agencies, art galleries and museums.

HYATT® HOTELS

©1983 Hyatt Hotels Corp.



HIGHLAND BASIN leading to the cliff-edge jump-off at Head-Smashed-In covers an area of 36 square kilometers where the valleys of six streams tributary to Olsen Creek provided natural runways (*arrows*) for the stampeding buffalo. The rows of cairns helped the stampeders to keep the buffalo aimed toward their jump point (*color*), which was 20 meters above the talus slope in the second century A.D.

dian period, before the Altithermal began, to 200 years after present climatic conditions had returned to the prairies.

The geologic note may surprise those who know the Great Plains as a generally featureless flatland. So they are, flat as a pancake if one neglects the terrain associated with stream erosion, occasional hilly sections and the prairie mountain ranges. The flatness, however, is the product of the Pleistocene glaciation that mantled the preglacial landscape with a thick cover of glacial till. The till is virtually a pudding of potential raw material for stone tools: quantities of quartzite and chert cobbles and pebbles. Moreover, at least in southern Alberta, the Cretaceous bedrock that underlies the till contains quartzite and chert conglomerates, moderate amounts of petrified wood and in one formation (the Bearspaw shale) quantities of black chert pebbles. Flint quarries in the Knife River area of North Dakota are another well-known source of tool material. They yield a honey-colored flint that was much sought after and widely traded.

The Rockies that form the western boundary of the prairies offer both similar and quite different kinds of tool material. High-quality cherts are common, but so are beds of metamorphosed siltstone and nodules of chalcedony. The most distinctive tool materials of all from the region are obsidians. Most of them came from the area that is now Yellowstone National Park, and Yellowstone obsidian was traded to as far away as sites of the Hopewell culture east of the Mississippi.

So much for the background. To return to the people of the Mummy Cave phase at Head-Smashed-In, they used local quartzite as the raw material for their projectile points. They prepared stone cores from quartzite cobbles, struck blanks from the cores and reworked the blanks into projectile points that were bifacial (worked on both sides) and notched at the sides. The size of the points suggests they were mounted on the kind of dart shaft that was propelled by a hand-held throwing stick (called by the Aztecs an atlatl). The Mummy Cave people also struck large, irregularly shaped flakes from quartzite cobbles, but evidently they fashioned very few of the skinning knives characteristic of succeeding cultures; we found only one. They were also not much interested in exotic tool materials. Only 1 percent of their points were made of imported stone: chert from Montana.

Overlying the four strata left by the Mummy Cave people was a sterile layer of talus wash and windblown sand that ranged in thickness from a little more than half a meter to nearly two meters. A mature grassland soil had developed on the top of this layer, suggesting that the buffalo jump had not been in active use for some centuries. Above the sterile layer were a series of bone beds with a total thickness of one to two meters. With one anomalous exception, the carbon-14 dates for these strata ranged from 1090 ± 120 B.C. at the bottom to A.D. 25 ± 80 at the top. All four strata represented a well-known buffalo-hunting culture of the northern plains and Rockies known as Pelican Lake after a site in southern Saskatchewan.

Among the buffalo bones were characteristic Pelican Lake projectile points, which are notched at the corners. They had been made from bladelike flakes struck off conical or polyhedral cores of good-quality chert. The Pelican Lake

COLORTRAK



PALE BY COMPARISON.

That's how a conventional 19-inch TV measures up against this ColorTrak 2000. RCA's remarkable new ColorTrak 2000 model #FJR2020W puts 25" of picture (measured diagonally) into the same space as conventional 19" sets. And along with 70% more picture, this new ColorTrak comes with a combination of exclusive RCA features: 58-function Digital Command Center remote control, high-compliance speakers, 127-channel tuning (including cable) and RCA's advanced detail processor, which delivers a picture that appears almost 3-dimensional. See your RCA dealer for a demonstration. And see 25" of picture that sits in 19" of set. Beautifully.

For more information and a free copy of the "Living with Video" book ($2.50 retail value), write: RCA Consumer Electronics, Department 32-312L, P.O. Box 1976, Indianapolis, Indiana 46206.

WE'LL OPEN YOUR EYES.

RCA

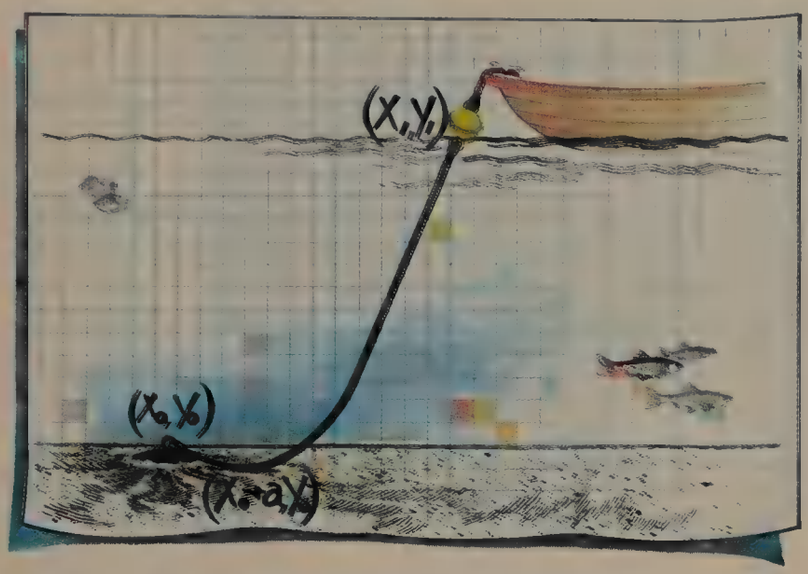
TWO THOUSAND

The Texas Instruments
Because y(x)=(1/k) co



Home Computer.
sh [k(x+a)]+b.

Whether it's beginning math or this catenary equation, the TI-99/4A Home Computer gives you real power to solve real problems.



The problem: A 100-foot-long cable, whose weight is 0.67 pounds/foot (corrected for buoyancy) is used to secure a mooring buoy to an anchor in 30 feet of water in a flat-bottomed lake. Determine the horizontal and vertical force components at the anchor and buoy when the buoy is displaced horizontally 92 feet from the anchor. If any of the cable lies on the bottom of the lake, find the distance from the anchor at which it breaks away from the bottom. When the cable is unconstrained (i.e., not touching bottom), the equation of the catenary is $y(x) = (1/k)\cosh[k(x+a)]+b$. See the computer screen below for the solution.

A true home computer should be user friendly to every family member, novice or professional. It takes power to do it.

That's what makes the TI-99/4A Home Computer the one to choose. It's a powerful package designed around our 16-bit TMS 9900 microprocessor. And it contains a lot of fully usable memory – 16K RAM and 26K ROM, plus up to 38K more ROM in each solid state cartridge.

With TI's BASIC system built into ROM, it doesn't use up Random Access Memory (RAM). Others, who load their operating systems into RAM, can use up half their memory before you even get started.

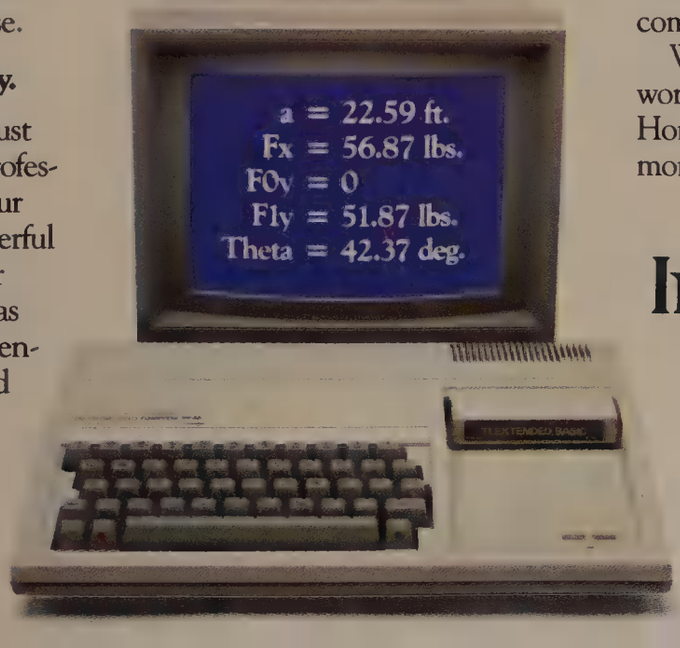
The result is that the TI Home Computer performs highly sophisticated computing tasks as easily as it teaches third grade math. For the money, no other computer is better suited for technical or scientific use.

Greater arithmetic accuracy.

Our unique 16-bit processor is just like the one found in expensive professional computers. Coupled with our rich TI BASIC, or still more powerful TI Extended BASIC, it makes our scientific programming capability as good as that provided by many scientific languages for mainframes, and better than some.

We give you 13-digit accuracy, and number range spanning $1x10^{\pm 127}$. All other home computers, using 8-bit processors, limit you to 6 to 9 digits, and only $1x10^{\pm 38}$ range. It's a big difference. And it means you can solve problems on the TI Home Computer which aren't feasible on the others.

TI BASIC also gives you more powerful commands and better flexibility for naming variables, since it recognizes up to 15 character descriptive names. Some others recognize only the first two significant characters in names, which means they can't tell the difference between "TAX" and "TARIFFS." So you have to resort to confusing acronyms.



The TI Home Computer is fluent in seven languages.

You can choose languages for all programming levels, from built-in TI BASIC to TI Extended BASIC, TI LOGO II, UCSD Pascal,* TI PILOT, TI FORTH and our powerful TMS 9900 Assembly Language.

You can also write useful non-scientific programs with the TI Home Computer. It's the only one in its price range that lets you combine 16-color graphics, single-command animation, sound effects, music, and speech with unlimited vocabulary into the same TI BASIC program.

And only TI's home computer comes with a 1-year warranty.

Whatever kind of problem you're working on, work it out with the TI Home Computer. You can't find more computer for the money.

TEXAS INSTRUMENTS

Creating useful products and services for you.

*UCSD Pascal is a trademark of the Regents of the University of California.

Copyright © 1983 Texas Instruments



SANDSTONE CLIFF at Head-Smashed-In stands about 10 meters above the talus slope at the southeastern edge of the Porcupine Hills, a region of rolling terrain. The Plains Indians of the culture named Mummy Cave may have been the first to drive buffalo herds over the cliff to their death on the slope below. Three other groups of buffalo hunters, including the predecessors of the Piegan Indians, who gave the site its name, used the jump thereafter. The trenching on the slope, to the right of center, was done by the author and his associates in 1965 when the Glenbow-Alberta Institute and the University of Calgary began an investigation of the site.



MAIN TEST AREA was a three-by-six-meter shaft, to the right of center, that was pinched down to a narrower shaft at a depth of eight meters. The material from the shaft, which eventually reached a depth of 11 meters, was brought to the surface by means of the improvised crane seen in action here and screened for its contents. Buffalo bones from the eight-meter level, when they were dated by carbon-14 analysis, proved to be about 3,000 years old. Objects from the bottom of the lower shaft, however, were too few to yield a culture identity or a date.

people, unlike their predecessors, had a taste for exotic tool materials. Their knives (large bifaces), their scrapers and their smaller retouched flake tools were made from Rocky Mountain cherts and chalcedonies, Yellowstone obsidian and the honey-colored flint from the Knife River quarries in North Dakota.

Some of the projectile points from the Pelican Lake strata have stems rather than being notched at the corners. These too were made from an exotic raw material: metamorphosed argillite from the Kootenay Lake area of British Columbia some 800 kilometers west of Head-Smashed-In. Both their source and their alien shape suggest that visitors from Kootenay Lake occasionally participated in the Pelican Lake people's buffalo drives.

One attribute the Pelican Lake people had in common with their predecessors was the use of the atlatl in hunting. Both whole and broken atlatl projectile points were found among the buffalo bones. The bases of broken points, however, were uncommon, suggesting that atlatl dart shafts with projectile tips broken off were usually retrieved and taken back to camp for rehafting. Such retrieval makes sense; an atlatl dart shaft, larger than an arrow shaft and probably made with far greater care, is the critical element in the atlatl ballistic system, just as the bow is in the bow-and-arrow system. As will be seen, once bows were substituted for atlatls, the bowman hunters considered arrows and their arrowheads expendable.

The second cliff-face slump took place in about A.D. 100. In the southern section of the talus massive chunks of sandstone buried themselves in the uppermost Pelican Lake stratum. To the north an entire section of cliff, ranging from 15 to 21 meters in thickness, fell a distance of 12 meters, rotated and then slid down along the tilted bedrock, pushing some of the overlying Pelican Lake strata downhill ahead of it. Although buffalo driving appears not to have been seriously interrupted by the catastrophe, the northern half of the talus slope was little used thereafter. The initial impact of the slump block had left a deep depression; the hunters exploited it as a natural corral, driving the buffalo over the cliff so that most of the dead and maimed animals ended up in the hollow. Among the buffalo bones that had accumulated there we encountered arrowheads for the first time.

The distinctive small points for arrows usher in the Avonlea phase at Head-Smashed-In. The phase takes its name from a site in Saskatchewan, and the Avonlea people were sophisticated buffalo hunters, equally adept at driving the animals over a precipice or herding them into a corral. The culture may have evolved from the Pelican Lake

Buick's leather-bound classic.

Space permitting, we could write a book about the 1984 Riviera: the independent rear suspension, the front-wheel-drive technology, the computer-monitored chassis, and much more.

But in spite of all the rational, high-tech reasons to admire Riviera, there's another, more time-honored factor.

Because Riviera is a classic. With the unmistakable flair of top-down motoring in genuine hand-fitted leather-trimmed seating.

And, of course, items such as air conditioning, power seats, and automatic level control are standard.

Coupe or Convertible, buckle yourself into a genuine classic, at your Buick dealer.

Official Car of the XXIIIrd Olympiad
Los Angeles 1984

BUICK

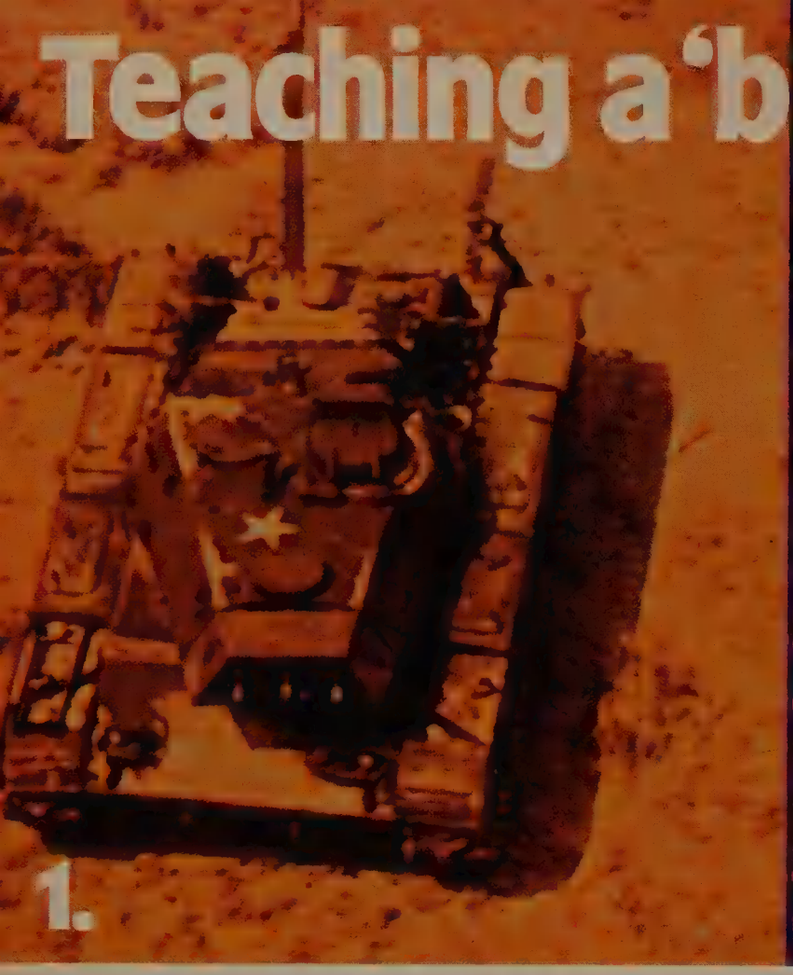
Wouldn't you really rather have a Buick?



Some Buicks are equipped with engines produced by other GM divisions, subsidiaries or affiliated companies worldwide. See your Buick dealer for details.

GM

Teaching a 'blind' computer





Lockheed knows how.

One of the growth technologies of the future is computerized image processing.

It's a vital technological thrust. More and more, industry, medicine, science and defense are making use of huge numbers of images gained from many sources—radar, infrared, acoustic, television cameras.

Until recently, most images have been analyzed or compared only by specialists. But now, computers are being taught to "read/understand" and make decisions about torrents of images that would otherwise swamp the human eye and brain.

"The tanks just moved."



The orange lines, which are offset from the original image of the tank, are part of the recognition technique that shows the tank has moved.

This can be critical input to a threatened commander, and he needs it almost instantly.

A reconnaissance aircraft or satellite carrying a sensor such as a television camera or radar would make the discovery. The sensor's images then are converted into electrical signals and sent to a computer, where they are stored as arrays of numbers. Computer programs manipulate the numbers until the tanks are "recognized." An enhanced image showing the tanks can be reconstructed and displayed for the commander.

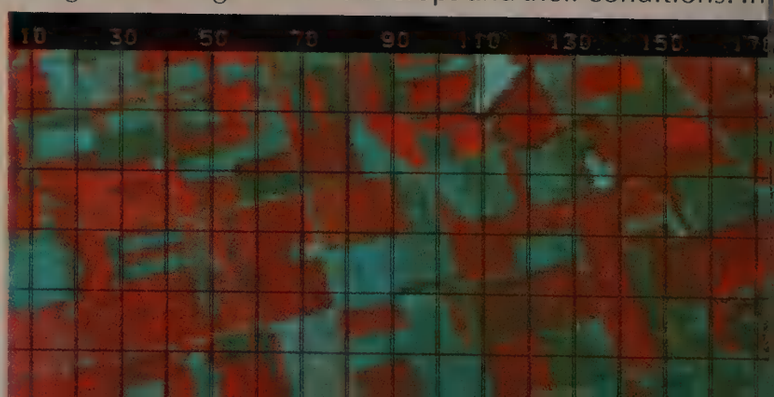
But a computer, unlike a human being, has to be given a mathematical description in advance to know what a tank looks like. The sensor, covering everything it sees, sends a staggering volume of signals. So the computer must be "trained" to eliminate unimportant images and concentrate on significant ones . . . for example, to look for the tanks that were spotted yesterday to see if they moved.

This is not just laboratory theory. Lockheed teams right now are training computers in this automated image understanding, or "computer vision."

How big is the wheat crop?

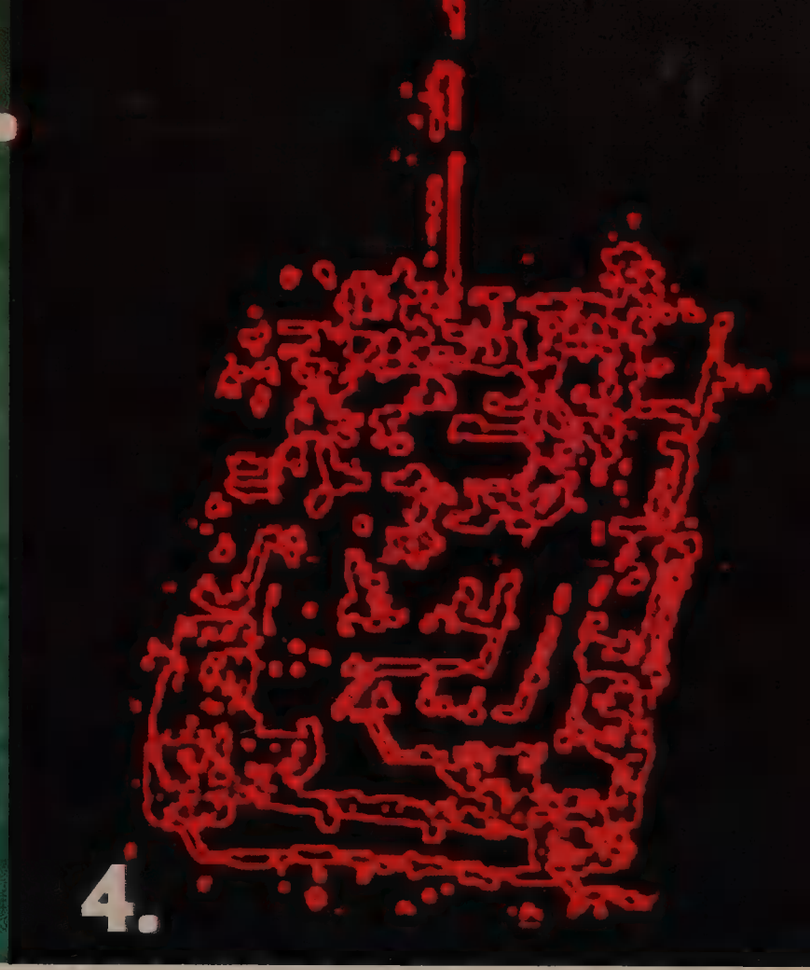
Identifying wheat from corn from sagebrush or other growth is another triumph of computerized image processin

The photo here came from a Landsat satellite. The red color, assigned by a computer, indicates wheat. Bare ground, plowed ground, and stubble are in shades of green. Working with NASA, Lockheed scientists, engineers and mathematicians have developed techniques to process hug numbers of images through computers, which they have "taught" to recognize various crops and their conditions. In



to 'see' a tank.





the human eye, a tank looks like the picture in Frame 1. But to a computer, the picture is simply a large array of numbers stored in its memory. It has to cess that data to "recognize" the tank. In Frame 2, it analyzes light intensities of the picture. In Frame 3, it manipulates its stored data to identify edges of tank. And in Frame 4, it has eliminated background and now has abstract edges and shapes from which it can proceed to recognize the tank.

e today, this valuable technology enables Lockheed to termine the size of world crops.

avigating by images.

How do you navigate an unmanned vehicle when an ertial guidance system is too expensive?

One way is with an image-based system being developed Lockheed.

Like a human pilot, the computerized system will "read" mbers of cues, like woods, ridges, and hills. It will recog- e a lake, for instance, even when iced over or shrunken drought. With this input, along with other data such as speed and compass readings, the computer will weigh its information and initiate accurate navigational orders.

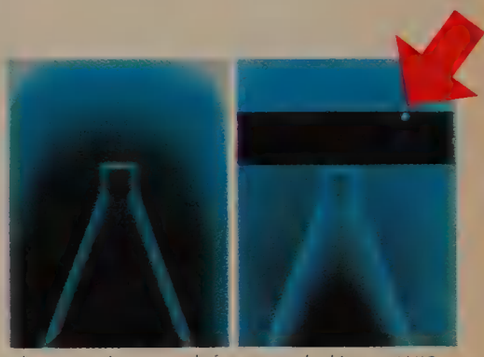
Lacking human eyes that can tell a lake from a hill, the mputer will rely on mathematical descriptions of topo- phic features. These descriptions, involving such features edges or brightness contrasts, will "teach" the computer to d and react correctly to the three-dimensional world.

utomating X-ray inspection.

Detecting subtle differences or flaws shown by X-ray otographs has traditionally been a slow, tiring job for man specialists, particularly in industry.

But now Lockheed has developed an Automatic X-ray pection System—AXIS—that examines X-rays by the thou- nds, at great speed. It converts X-ray images into numbers and mpares the results against established norms, thus isolating fects. With quick, tireless repeatability, it makes accept/ ect decisions and records the data on magnetic tape.

First developed for the defense industry, AXIS can be easily adapted to medicine and other industries. It will be increasingly important in applications where large numbers of images or articles must be checked fast for subtle, critical variations.



Arrow points to a defect revealed in an AXIS X-ray image of an artillery shell casing.

Image processing: the future.

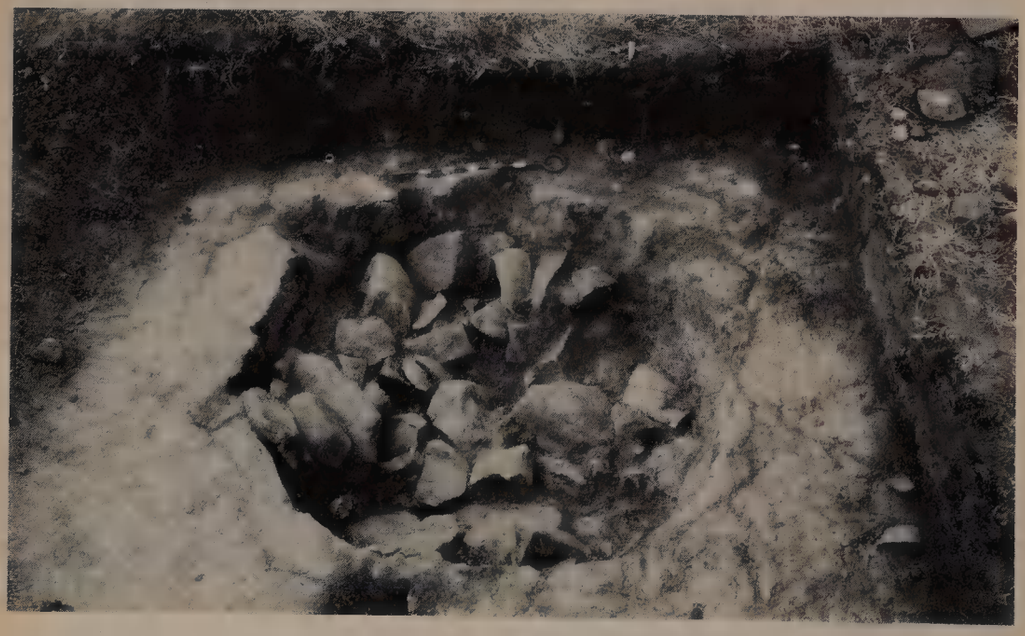
If a system of sensors and computers can recognize instantly the movement of tanks or scan the ground below to guide an unmanned vehicle, cannot more advanced systems be developed to guide surgeons as they operate? Cannot systems such as AXIS lead to immense improvement in the quality of American products?

Great as have been the advances in image processing, it is a field only now coming into its own, and Lockheed knows how to develop this emerging technology.

Lockheed



EXCAVATOR WITH TROWEL works to expose additional buffalo bones in the east area of Head-Smashed-In during the 1972 season. Jaw fragment and rib cage have been cleared.



PROCESSING PIT in the campsite adjacent to Head-Smashed-In was used to boil the fat out of buffalo marrowbones. A buffalo hide, lining the pit, was filled with water heated by dropping into it stones that had been baked in a campfire. The fat, ladled out of the pit and allowed to congeal, was the base for pemmican: a mixture of foodstuffs that could keep for months.

complex itself. At least the first Avonlea arrowheads at Head-Smashed-In are corner-notched ones made of chert obtained from the Pelican Lake quarries in Saskatchewan. The evolution, if such it was, did not take place at Head-Smashed-In and its vicinity. Of all the points we recovered from the lowest Avonlea strata, points for atlatl darts make up less than 1 percent of the total.

The Avonlea people are the first among the plains populations to have adopted the bow and arrow for hunting. They acquired the new weapon through contact with fishing people in the interior of British Columbia, who in turn had learned about bows and arrows from Indians of the Northwest Coast whose bowmanship was by then 1,000 years old. At Head-Smashed-In the Avonlea people drove buffalo in such numbers over the cliff edge that the natural corral left behind by the slump came to hold a thick bed of bones. In the lowest part of the three- to four-meter Avonlea deposit, which is dated at about A.D. 110 ± 90, Avonlea arrowheads were numerous. Many of them were unbroken; others were represented only by the tip of the head or by the base of the head and its haft. Evidently the Avonlea bowmen made no particular effort to retrieve their arrows after each stampede, perhaps because many arrow shafts would be broken when the falling buffalo hit the ground.

The Avonlea people imported very little Yellowstone obsidian and no chert at all from the Montana mountain quarries as raw material for their arrowheads and other stone artifacts. Instead they brought chert of exceptionally high quality in from British Columbia and from other sources that are as yet unidentified. In any event, working with imported chert, the Avonlea craftsmen turned out the finest stonework seen on the Great Plains since the time of the Paleo-Indians, some 8,000 years earlier. Their arrowheads, made from small blades struck from conical cores, are extremely thin and were delicately flaked into final shape. Other distinctive Avonlea tools included skinning knives, end scrapers, small flakes and wedges, the latter perhaps serving to split buffalo leg bones for their marrow. At first the exploitation of local stone was confined to heavy-duty implements for chopping, hacking and smashing, but as time passed the Avonlea toolmakers came to depend more and more on Great Plains stone.

The upper Avonlea levels, which are dated at about A.D. 850, differ in character from the lower ones. Here the massive lens-shaped deposits of buffalo bone are heavily charred and in some instances completely calcined. Above them lie a final series of bone beds, 2.5 meters deep, associated with the last

When it comes to superior performance, we study our lines very carefully.

Superior printer performance is not a fluke. It evolves from analyzing printed line after printed line. Taking the time to test and retest. After 30 years of manufacturing precision parts, we know that there are no shortcuts.

And so we took the Gemini-10X and methodically put it through its 120 cps pace. We achieved a print head life of over 100 million characters with an extremely precise dot alignment creating each crisp character.

So far so good.

Next, sophisticated performance demanded versatility. A wide choice of character sets, a buffer expandable to 8K, and the ability to interface with all popular personal computers. We added macro instruction, giving Gemini-10X the capability to perform up to 16 operations with one command. We included as standard a paper feed system that has a friction and fully adjustable tractor feed. Then we even built in the dexterity to print graphics and text on the same line.

Done.

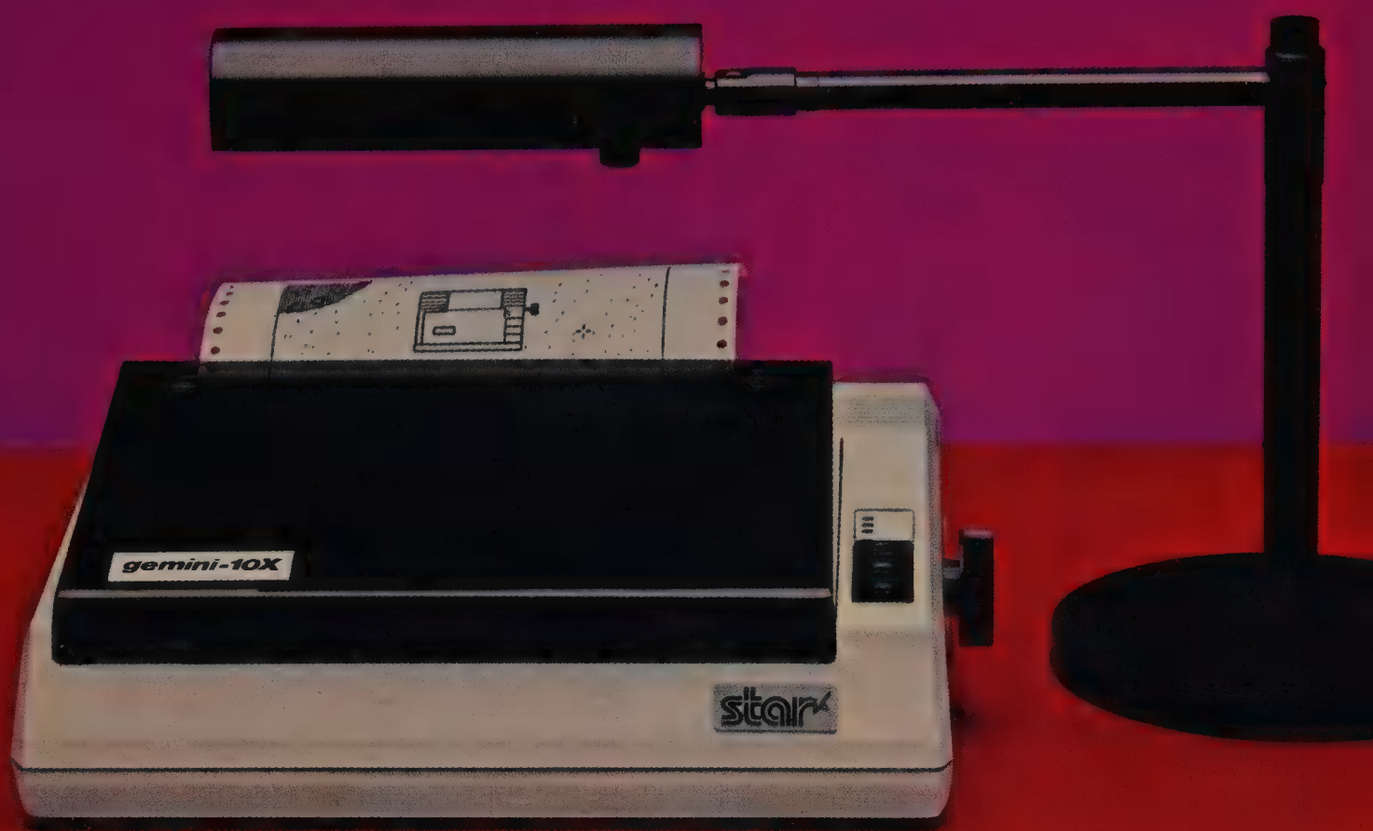
And, of course, staying the best means constant reviewing and fine-tuning. Keeping the Gemini easy to find, easy to afford and so reliable it can be warranted for up to twice as long as its major competitors.

Only the most careful engineering has built the new hard-working Gemini-10X. You'll applaud its performance.

star™ micronics·inc

THE POWER BEHIND THE PRINTED WORD.

Computer Peripherals Division
2803 N.W. 12th Street, Dallas/Ft. Worth Airport, TX 75261



ENERGY OPTIONS
The Practical and The Promising



The energy crises of the 1970's jolted America into greater awareness of alternative energy development. Some of the alternative sources being explored today will play an increasing role in the years to come. The question is, how much will the new alternatives be able to supplement our major energy sources?

Although overall energy demand in the U.S. dropped slightly over the last decade, the use of electricity actually rose more than 20 percent. This electrification trend has gone on throughout the century and is expected to continue. The United States Department of Energy recently reported that "the economy is very likely to turn increasingly to electricity for its energy requirements."

As this reliance on electric energy continues to grow, *all* the resources that can generate electricity are being examined. Some of the alternatives are discussed here. Others, like fusion and ocean thermal energy conversion, may play a role in the more distant future.

Sun, wind, earth, and water

SOLAR. In some respects, solar energy has begun to pay off. For the heating of water, solar systems are now economically competitive in some regions of the U.S. with plenty of sunshine. Also, there are new buildings benefiting from "passive" solar designs. They utilize the sun through nonmechanical means such as skylights and heat-storing rock beds.

For the generation of electricity, however, solar technologies are in early stages of development. Although some limited uses are being made, solar electricity is still too expensive for widespread practical use.

One possible system is the solar-thermal electric station, or "power tower." A small solar-thermal plant in the Mojave Desert uses a large field of mirrors to track the sun and focus its heat onto a centrally mounted steam boiler that drives a turbine-generator.

Another solar electric technology is photovoltaics. It uses silicon or other semiconductors to convert the sun's rays directly into electricity. Researchers are working on increasing the efficiency and reducing the high costs of photovoltaics.

WIND. Wind power, like many alternative energy sources, depends on geography. It is most promising in the Great Plains, in mountain passes,

and along coastal areas, where winds are strong and steady.

Some of the larger wind projects have had problems. But they are being addressed, and utility companies today are involved in well over 100 wind-power research projects. California's Energy Commission has



Reliability and strength of the wind play an important role in the location of "wind farms" like this one in Altamont Pass, east of San Francisco. Wind generators are spread over hundreds of acres of land.

predicted a wind-power potential of 1000 megawatts for the state by the year 2000, if current trends continue. That's the equivalent of one major electric power plant, or approximately 2 percent of the state's present electrical capacity.

GEOTHERMAL. The internal heat of the earth is another energy source. Geothermal energy is captured from hot water or steam created by hot or molten rock. 1300 megawatts of geothermal electric capacity are now available in this country, mostly in California.

HYDRO. Falling water generated almost 14 percent of the nation's electricity last year. Most U.S. sites for large-scale hydro projects are already in use, but there is potential for further development of low-head or "minihydro": the use of small and medium-sized dams, either newly built or upgraded. By 1985, there may be about 400 new small-scale hydro projects, largely in the Northeast. Like other energy sources, though, hydro has limitations. These include the environmental changes that can be caused by damming a river, as well as the uncertain reliability of power that is fed by rain and snow.

BIOMASS. This term refers to a wide variety of fuel sources, including plant and animal waste, wood, algae, and garbage. The largest user of biomass in the U.S. today is the forest products industry, which gets about 50 percent of its energy from wood wastes. Another form of biomass is methane gas produced from decaying organic materials.

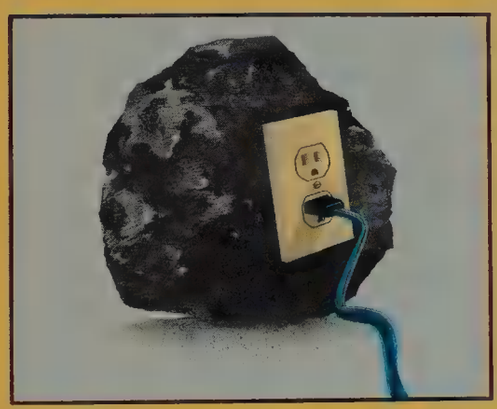
Energy sources we can count on today

Some of our important energy options involve not new production technologies but more efficient ways of delivering and using energy. So conservation methods, including industrial cogeneration, will also help meet our energy needs.

With all of America's energy alternatives, it's easy to forget how much of our electric power will still be generated from conventional fuels. The Department of Energy projects that by the year 2000, over 80 percent of the country's total electricity generation will still come from the abundant domestic resources of coal and uranium.

Oil and natural gas are more valuable fuels for transportation and manufacturing than for burning in power plants. Also, these fuels, especially oil, are subject to uncertainties over price and future supply.

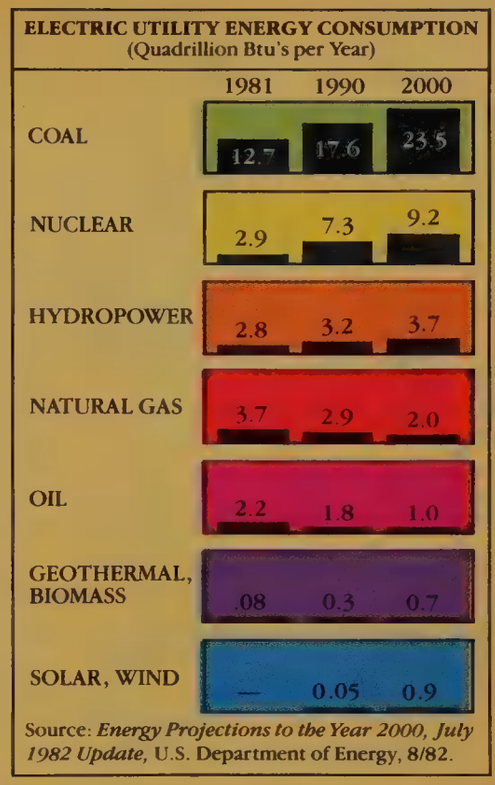
On the other hand, we have enough coal to last hundreds of years. Coal will remain our dominant



We have enough coal in the U.S. to last hundreds of years. It is estimated that by 1990, electric utilities will be burning over 780 million tons of coal a year to supply more than half of America's electricity.

source of electricity. And advanced technologies such as coal gasification and fluidized-bed combustion are creating new ways to burn coal cleanly.

To secure a well-balanced supply of energy, America is also relying on nuclear power. Aside from the 80 plants already generating nuclear electricity in the U.S., 56 more have been granted construction permits by



The Department of Energy foresees sizable growth for both nuclear energy and coal in this country's electrical future.

the U.S. Nuclear Regulatory Commission. And they are being built.

A steady and secure supply of electricity is critical for a healthy economy. Some of the alternative energy sources discussed here may eventually become significant contributors to America's electrical supply, depending on what part of the country you live in. Some may not. Whatever the future holds, we need to take advantage of the domestic energy resources already within our reach. We need to understand not only the electricity potential of tomorrow, but the electricity realities of today.

For a free booklet on *Energy Options* and the electrical future of America, send this coupon to:

U.S. Committee for Energy Awareness
P.O. Box 37012
Washington, D.C. 20013

NAME _____

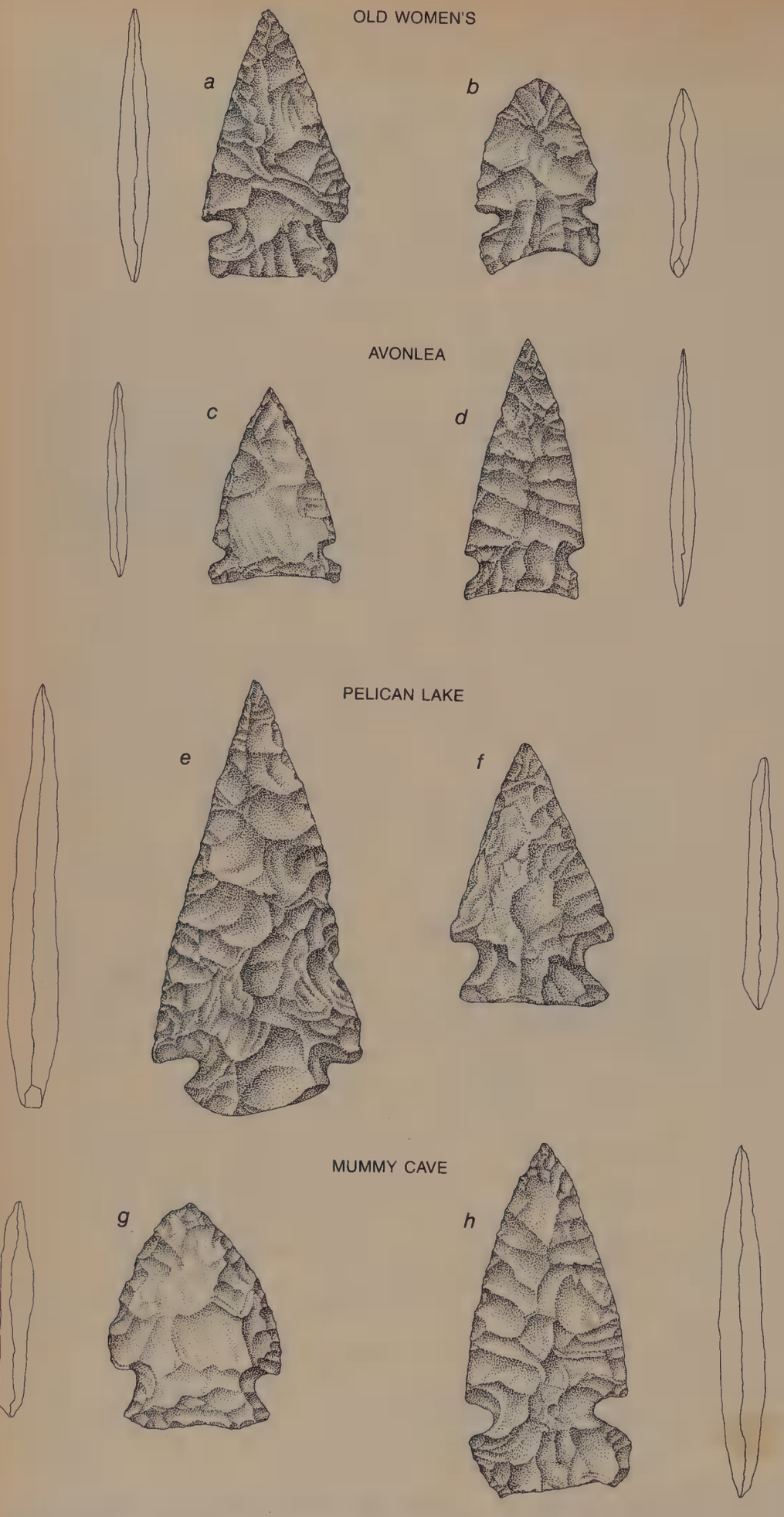
ADDRESS _____

CITY _____ STATE _____ ZIP _____

325

Information about energy America can count on today

U.S. COMMITTEE FOR ENERGY AWARENESS



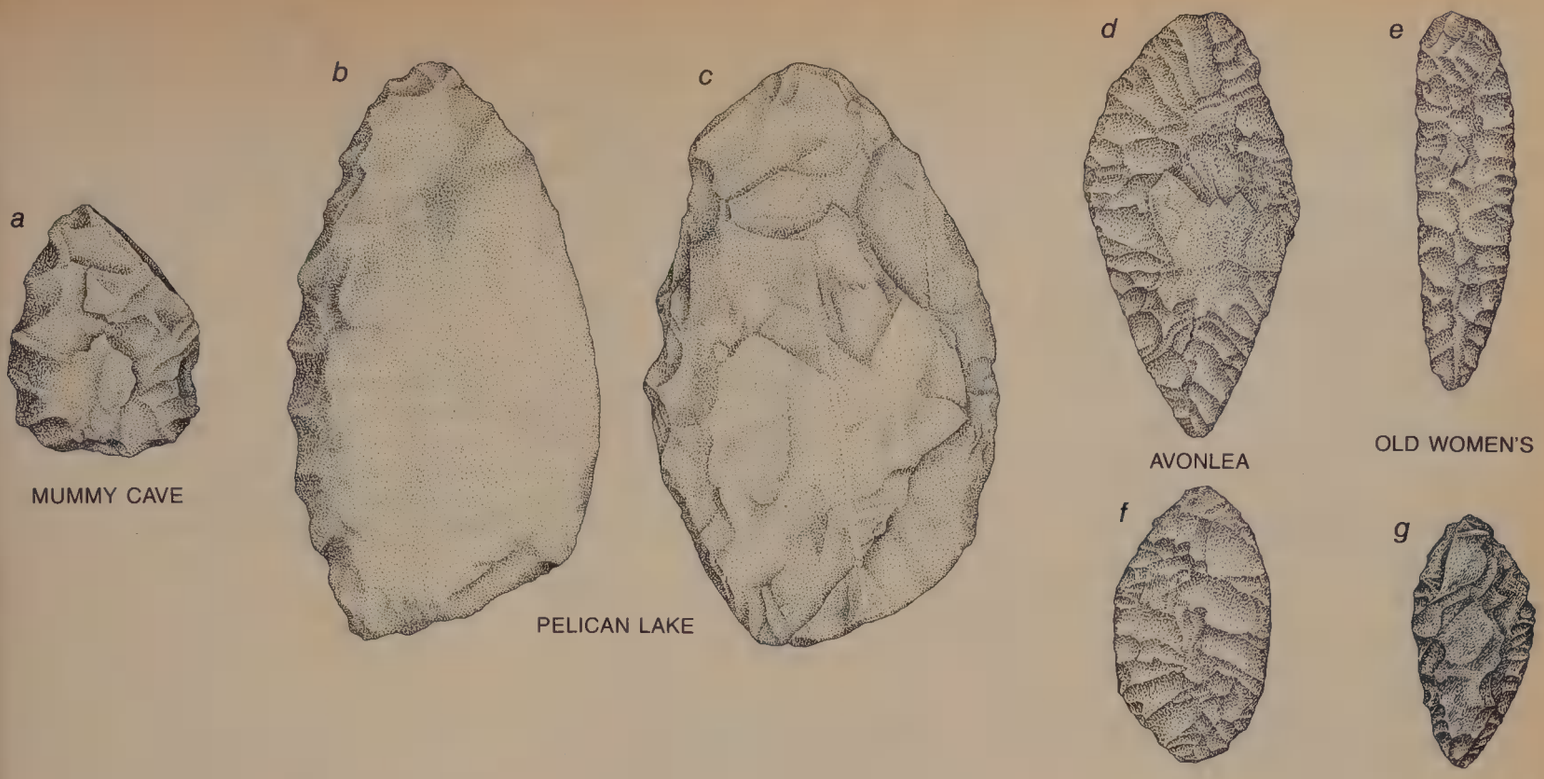
EIGHT PROJECTILE POINTS from the four phases uncovered at Head-Smashed-In appear in profile and edge on. The lower four (*e–h*) are points for atlatl darts; point *e* is obsidian. The upper four (*a–d*) are arrowheads; the atlatl, a dart-launching stick, was replaced by the bow at the start of the Avonlea phase at the site. The flaking done by the Avonlea people was the best on the plains following the end of the Paleo-Indian period. Point *d*, an obsidian piece, is a good example of Avonlea workmanship; it can be compared with point *a*, which, although it too is obsidian, has a much rougher finish. All the points are shown enlarged 50 percent.

culture phase at Head-Smashed-In: Old Women's. The name is that of the nearby site excavated in 1958. This phase extended from A.D. 850 to the historical present (A.D. 1800), or shortly after Peter Fidler's contact with the Piegan at Stimson Creek, 80 kilometers to the northwest. It is probable that the Old Women's people were prehistoric Piegan and that their culture was a further evolutionary stage of the preceding Avonlea phase. The transitional strata between the two show a gradual change in arrowhead style and in other stone technology.

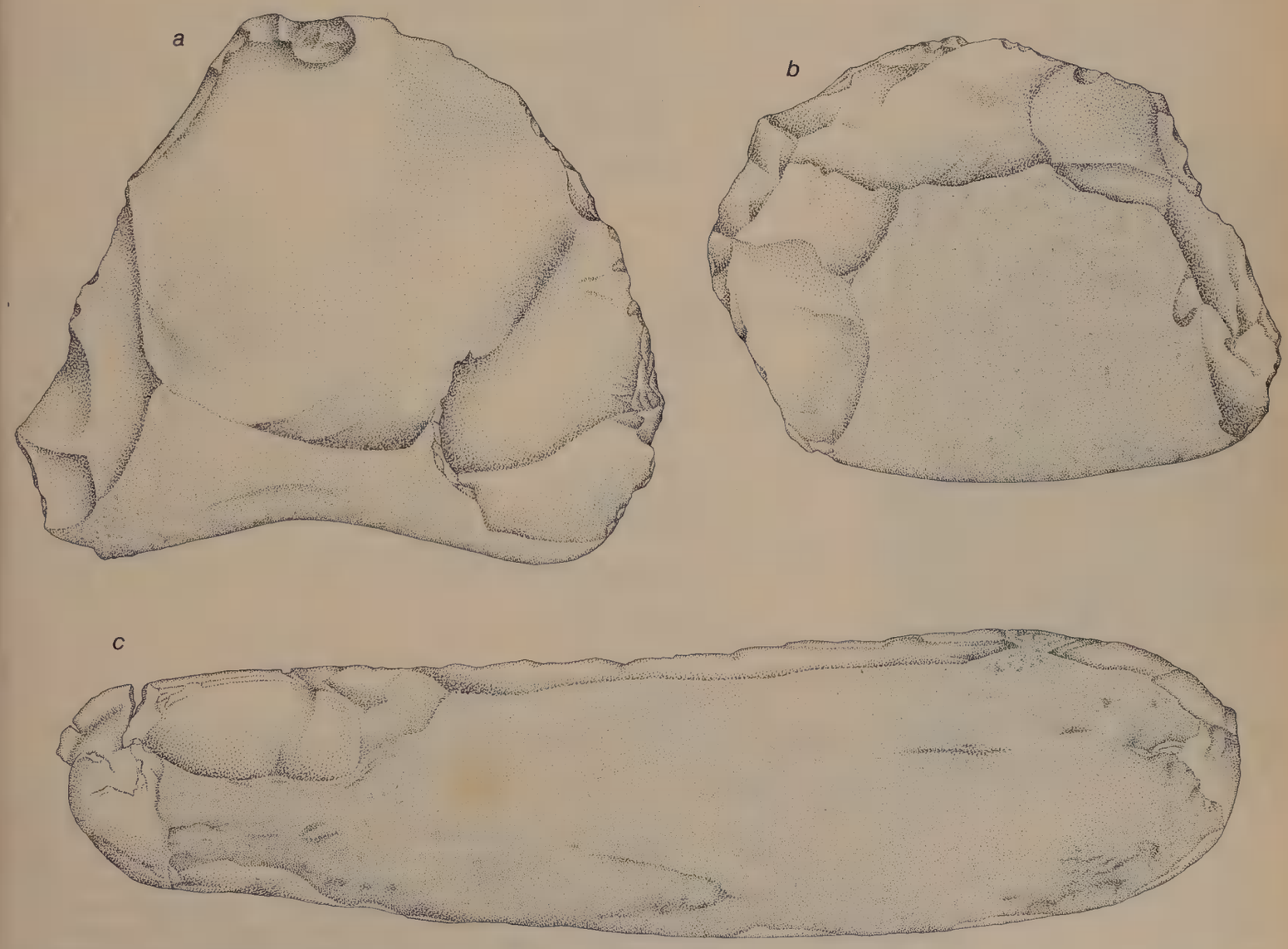
The Old Women's phase is characterized by small, side-notched arrowheads, cigar-shaped skinning knives hafted sideways into wood or bone handles, end scrapers and wedges, all made from local pebbles of chert. The Old Women's people, like their predecessors, also made use of large quartzite cobbles and flakes as choppers, bone-smashers, and anvils and hammers for striking flakes from small chert pebbles to make arrowheads and knives. Their workmanship in the early phase was not as fine as that of the Avonlea people, a fact that may be attributable at least in part to their reliance on local cherts of indifferent quality. They even made a few artifacts out of petrified wood. Near the end of the Old Women's phase, however, the toolmakers reduced their dependence on Great Plains raw materials and imported Yellowstone obsidian and porcellanite from southern Montana. A third exotic material, often used in earlier Head-Smashed-In phases, the distinctive honey-colored Knife River flint, was rarely exploited by the Old Women's craftsmen.

The buffalo-bone beds of the Old Women's phase varied considerably in composition. The one-meter strata of the lower deposits include lenses of charred bone among the unburned and well-preserved buffalo skeletons. Overlying this lower deposit was a stratum of charred and uncharred bone, buffalo horns and decomposed hair and hide, mingled with the contents of the animals' intestines, including feces. The final strata, 1.2 meters deep, contained well-preserved bones, unmarked by fire, in a powdery red orange matrix composed of decayed hide and hair. We found European trade goods—metal arrowheads—in the uppermost stratum and recovered a musket ball from its surface. On the basis of trade goods found here and at the nearby campsite we assigned to the upper part of the deposit a date of about A.D. 1800. The Old Women's deposit is topped by a layer of sandstone rubble, the remains of an unsuccessful 19th-century effort to establish a sandstone quarry at the site.

The campsite area, a broad, flat glacial bench 18 meters below the talus



SEVEN KNIVES representative of the phases at Head-Smashed-In are shown at four-fifths actual size. The number of knives from the site is low compared with the number of projectile points; the Mummy Cave strata included only one (*a*). Both the Avonlea and the Old Women's knives are more refined in flaking than the earlier ones. Old Women's knives were fitted into a wood handle along one edge.



HEAVY STONE TOOLS, made chiefly from quartzite found in the local glacial till, included flake choppers (*a*) and cobble choppers (*b*). The choppers were left mostly unmodified. Both kinds of tools were probably used more for breaking bones than for chopping meat. The evidence of wear on the clublike tool (*c*) suggests that it too was used to break bones. All three tools are shown at two-thirds actual size.

How to slim down. Rent a home. Retire. Choose a new carpet. Can food. Backpack. Buy tires. Find a job. Repair a leaky faucet. Read labels. Invest.

Questions? Let the Consumer Information Catalog help you find the answers. It lists over 200 helpful federal publications with answers to a whole lot of questions about your home, health, finances, and much more. For your free copy, write: Consumer Information Center, Department C, Pueblo, Colorado 81009.



Free.

General Services Administration
Consumer Information Center

YOUR EXCITING NEW HOBBY!

- Enjoy fantastic savings by assembling your own organ or piano.
- It's easy. No technical knowledge required.
- Just follow our clear, pictured instructions.
- Choose from many models from portables to consoles.
- Ask about our interest-free installment plan.

WERSI
ORGAN & PIANO KITS



For Free Information Pack M41
Call Toll Free
800-233-3865
In PA 717-299-4327
or send coupon below

WERSI Dept. M41 P.O. Box 5318
Lancaster, PA 17601

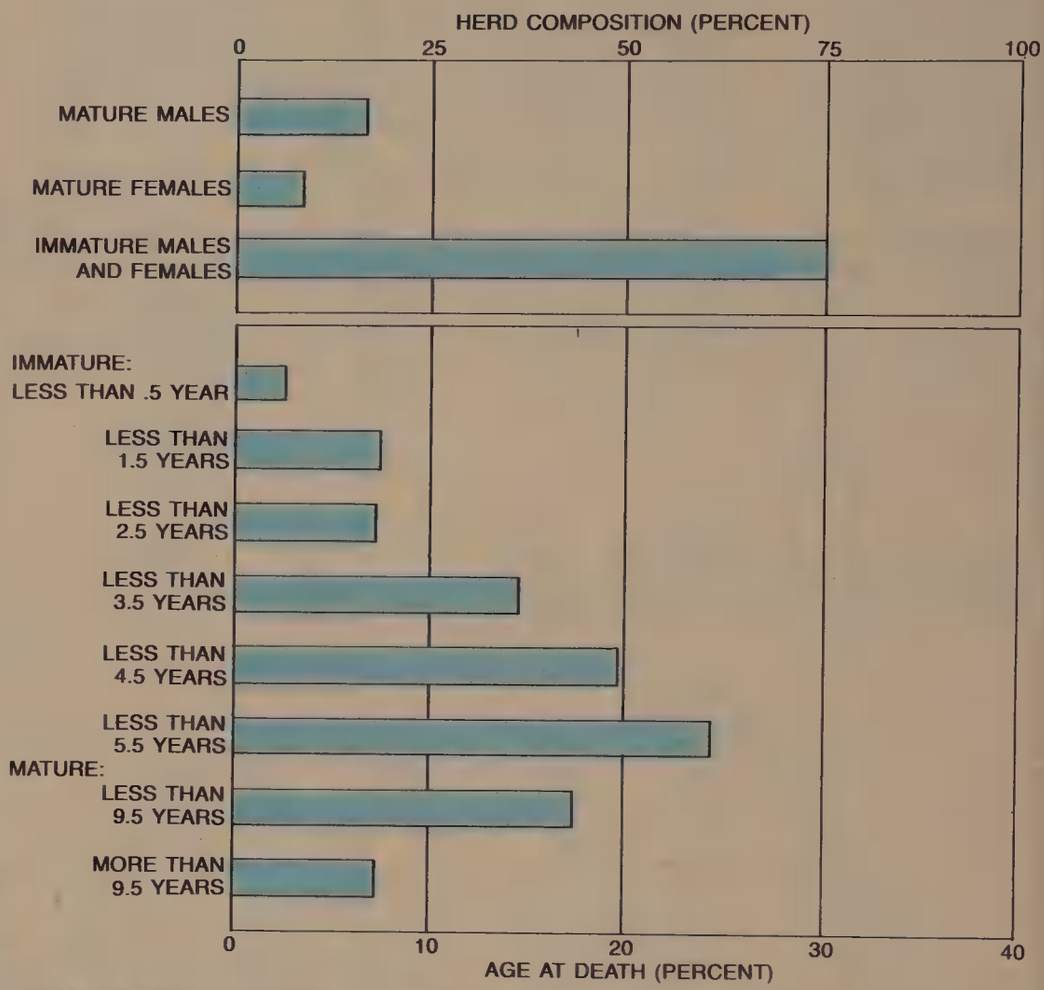
☐ Free Info. Pack: ☐ Organ ☐ Piano
☐ Catalog & Demo Record *(enclose $3)*

Name _____
Address _____
City _____ State _____ Zip _____
Phone (___) _____

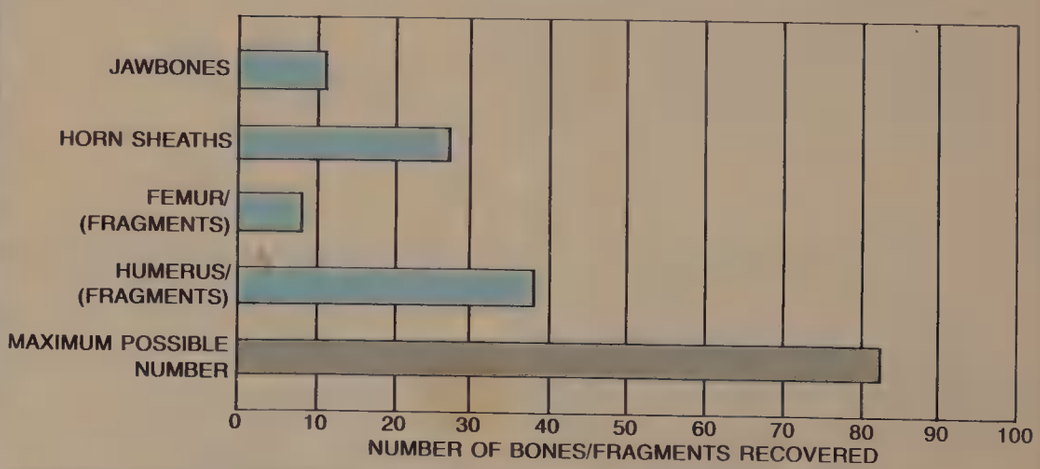
Rep inquiries invited

slope, is six and a half hectares in extent. Most of its present surface is mantled by a thin scatter of refuse: butchered bones, cracked rock hearths for cooking fires and miscellaneous artifacts. The preferred camping area was evidently the south end of the bench, upwind of the kill area. Here we uncovered a midden of bones, cracked rocks and artifacts associated with the final butchering of different parts of the slaughtered buffalo. Within the midden the principal artifacts were numerous heavy stone tools.

If historical accounts of Piegan buffa-



COMPOSITION OF BUFFALO HERD is based on an analysis of one bone bed at Head-Smashed-In (Avonlea North 2*B*), where the bones represented 41 individual buffalo. As the upper graph indicates, they were predominantly immature. The estimated age at death, as is shown in the lower graph, places seven of the animals in the fully mature category and only three in old age (more than 9.5 years old). Seven others were less than 2.5 years old and 24 were between 2.5 and 5.5 years old. Because buffalo cows usually calve in May, the animals must have been killed between four and five months later: from mid-September to late October.



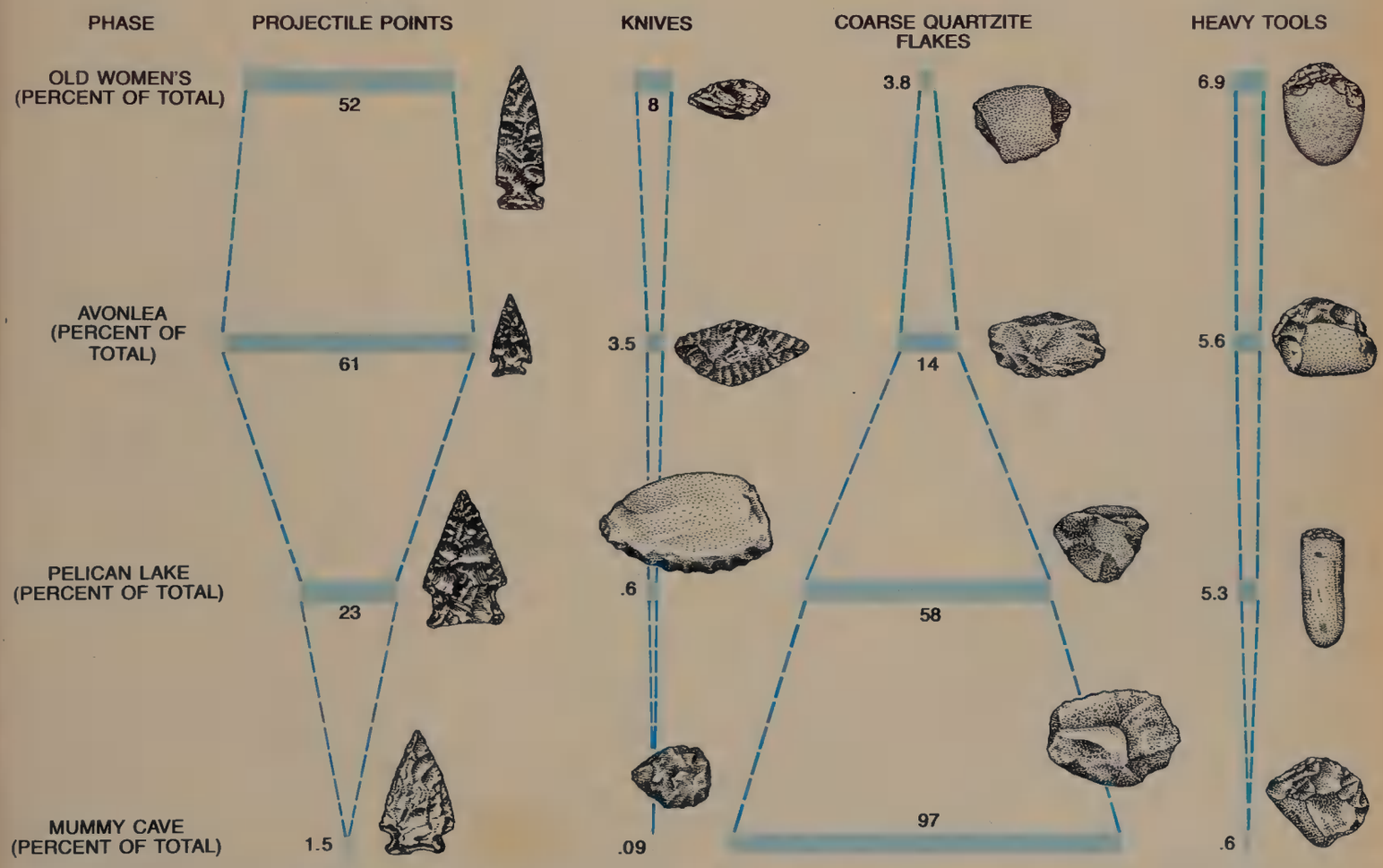
PRESERVATION OF BONES in the same Avonlea stratum, North 2*B*, shows that, as was the later practice, Avonlea hunters carried many parts of the slaughtered buffalo away from the kill site for further butchering. The presence of 41 individual animals was established by the analysis of teeth. If the carcasses had been undisturbed, 82 horns, 82 jawbones (upper and lower) and 82 front and rear upper-leg bones should have been found (*bottom bar*). Fewer than half that number of horns and only 11 jawbones were recovered. The number of femur (rear upper-leg bone) fragments was only eight. There were, however, 38 humerus (front upper-leg bone) fragments, suggesting that these bones were used less than femurs as a source of marrow.

lo hunts are any guide, the various stone tools had served to cut up the animals' four quarters, to cut the meat into strips for drying and to smash the leg bones for their marrow. Most of the stone cores and waste flakes at the camp had come from large, coarse-grained quartzite cobbles; little fine-grained chert waste was found. A small amount of pottery also turned up, along with such trade goods as beads, copper bangles, bottle glass and iron arrowheads.

Among the buffalo bones in the camp were fragments of mandibles, left over from the removal of the animals' tongue (considered a delicacy by buffalo hunters since Paleo-Indian times), the base and frontal bones of skulls (broken to remove the brain), rib bones, bones from buffalo humps and many fragments of limb bone. We also excavated a number of what, on the basis of their contents and of historical accounts, must have been boiling pits. Traditionally such pits were dug, up to two meters in diameter and one meter deep, and lined with a buffalo hide. The skin was then filled with water and cracked marrowbones. Rocks heated in a nearby fire were thrown in until the water boiled. Finally the marrow fat that accumulated on top of the water was skimmed off, ladled into containers and used as a base for pemmican: a mixture of the grease, berries and dried meat pounded to a powder.

Although the camp area may have been used by hunting parties other than those of the prehistoric and modern Piegan at Head-Smashed-In, we found little or no archaeological trace of any of the two earliest peoples who had stampeded buffalo at the site. Head-Smashed-In is, however, only the second buffalo-jump site to be excavated in Alberta, the first having been Old Women's. At least 20 more sites, with their associated hunting and butchering camps, still remain to be investigated. Our site provides a record of use by Great Plains hunters over a period of at least 5,600 years. That is a considerable span of time. Indeed, only one other excavated buffalo-jump site is substantially older: Bonfire Shelter, near the juncture of the Pecos River with the Rio Grande in Texas. Bonfire Shelter is some 10,000 years old, but the history of its use is far briefer than that of Head-Smashed-In. With the exception of Bonfire Shelter, jump sites are typically distributed along the western edge of the plains. Who knows how old some of the other Alberta buffalo jumps may be?

In any event our findings at Head-Smashed-In indicate that substantial numbers of hunters on foot were successful in providing themselves with meat and fat, with hides for robes and tepee covers, with bladders and lengths of intestine for containers, with sinew for sewing, with bone and horn for tools and utensils and with dung and tallow to burn for warmth and light. Moreover, their hunting methods required more sophisticated social mechanisms than many scholars have been prepared to credit even the historical plains Indians with. For all their deceptively simple material culture, these earliest plainsmen were at the same time excellent organizers and daring coordinators when many of the white man's ancestors were still simple hunters and gatherers in the chill forests of Mesolithic Europe.



STONE ARTIFACTS of different kinds varied in number from one phase to another at Head-Smashed-In. In the Mummy Cave strata, for example, of some 1,129 stone artifacts found, 97 percent were flakes of coarse-grained quartzite made from local quartzite cobbles. The Mummy Cave projectile points, 17 in all, were for hafting to the long darts thrown with the atlatl. Only one knife was found in the Mummy Cave strata. Of the 1,149 artifacts from the Pelican Lake strata, more than half were also flakes of local quartzite. Atlatl projectile points, some made from imported obsidian, accounted for another 23 percent of the total. A few knives were present, some also made from imported stone, while heavy stone tools such as choppers and mashers accounted for 5.3 percent of the total. In the Avonlea strata, arrowheads were the most abundant (61 percent of the total) among the 1,735 artifacts found. Knives were far more numerous than they were in the Pelican Lake phase: the total was 60. Avonlea craftsmen imported chert from British Columbia and obsidian from the Yellowstone area. The people of the final phase, Old Women's, left fewer arrowheads than their Avonlea predecessors had left but more knives. Up until historical times they worked mainly in local stone, using chert for points and knives and quartzite for heavy tools.

The Extinction of the Ammonites

Changes in the shells of these nautiluslike marine animals at the end of their long history suggest they were fighting a losing battle against more mobile, shell-crushing predators

by Peter Ward

At the end of the Cretaceous period of the Mesozoic era 65 million years ago many species of plants and animals, including many groups of marine invertebrates, suddenly died out. One of the most important groups of marine invertebrates to be extinguished was the ammonites. The ammonites, like their nearest living relatives the members of the genus *Nautilus,* were chambered cephalopods: mollusks that live in the outermost compartment of a shell they divide into chambers by secreting a series of walls behind them as they grow. Many investigators now believe the extinctions at the end of the Cretaceous period were caused by the direct or indirect effects of the collision of a large meteoritic body with the earth, a possibility first proposed by Luis W. Alvarez, Walter Alvarez and their co-workers at the University of California at Berkeley to account for the abnormally high levels of platinum-group metals found in sedimentary strata dating from this period [see "The Mass Extinctions of the Late Mesozoic," by Dale A. Russell; SCIENTIFIC AMERICAN, January, 1982]. The fossil record suggests, however, that the extinction of the ammonites was a consequence not of this catastrophe but of sweeping changes in the late Cretaceous marine ecosystem.

"Ammonites" is a term reserved for the Jurassic and Cretaceous representatives of the order *Ammonoidea.* The shells of the ammonites differ from those characteristic of the ammonoids throughout most of their 300-million-year history in ways that suggest these slow-moving and probably bottom-dwelling animals were encountering much stronger selective pressures than they had up to that time. The typical ammonoid shell was beautifully engineered for withstanding the pressure of seawater at the depths at which the animals lived but was poorly engineered for escaping or discouraging predators. Most ammonoid shells, like those of the modern nautilus, were planispiral (coiled in a plane), comparatively thin-walled and poorly streamlined. In the late Mesozoic many fast-swimming, shell-crushing marine predators arose, among them the modern bony fishes and many species of air-breathing marine reptiles. Their presence may explain the predominance of three types of shells among the Cretaceous ammonites: streamlined planispiral shells, heavily ornamented or armored planispiral shells and a wide variety of bizarrely shaped shells collectively called heteromorphs.

Much can be deduced about the agility, speed and mode of life of the new species from their shells alone, but a complete understanding of the defensive strategies to which the shell types bear silent witness requires some understanding of the biology of the animals living in them. For the most part only the shells and other hard parts of the ammonites have been preserved; rarely are fossils containing traces of organic material found. Thus the living fossil, the nautilus, has proved to be an invaluable source of information, allowing otherwise inexplicable or apparently maladaptive modifications of the ammonite shells to be explained.

The ultimate fate of the ammonites is still an enigma, in part because ammonite-bearing sediments encompassing the boundary between the Cretaceous period and the Tertiary have been found at only a few sites. In the late Cretaceous both the number of ammonites and the number of species of ammonites dropped sharply. What happened to the remaining, apparently resilient ammonite species is not clear, but studies of the fossils from the stratigraphic sections at Zumaya in Spain suggest they became extinct long before the proposed impact of the meteoritic body. Even if it is assumed that the impact hastened rather than caused the final extinction of the ammonites, however, there are loose ends to the story. Some of the late Cretaceous ammonite species closely resembled the nautiloids, the ancestors of the modern nautilus. Why should they have succumbed when the nautiloids survived?

The nautiloids, some species of which are found in strata dating from the Cambrian period of the Paleozoic era, are the oldest cephalopods. A very restricted group (that is, a group limited in numbers) called the bactritids developed within this subclass in the Silurian period of the Paleozoic; the bactritids may be the common ancestor of all the more recent cephalopods, including the modern cuttlefish, squid and octopus. The ammonoids apparently branched off from the bactritids in the early Devonian period of the Paleozoic some 395 million years ago.

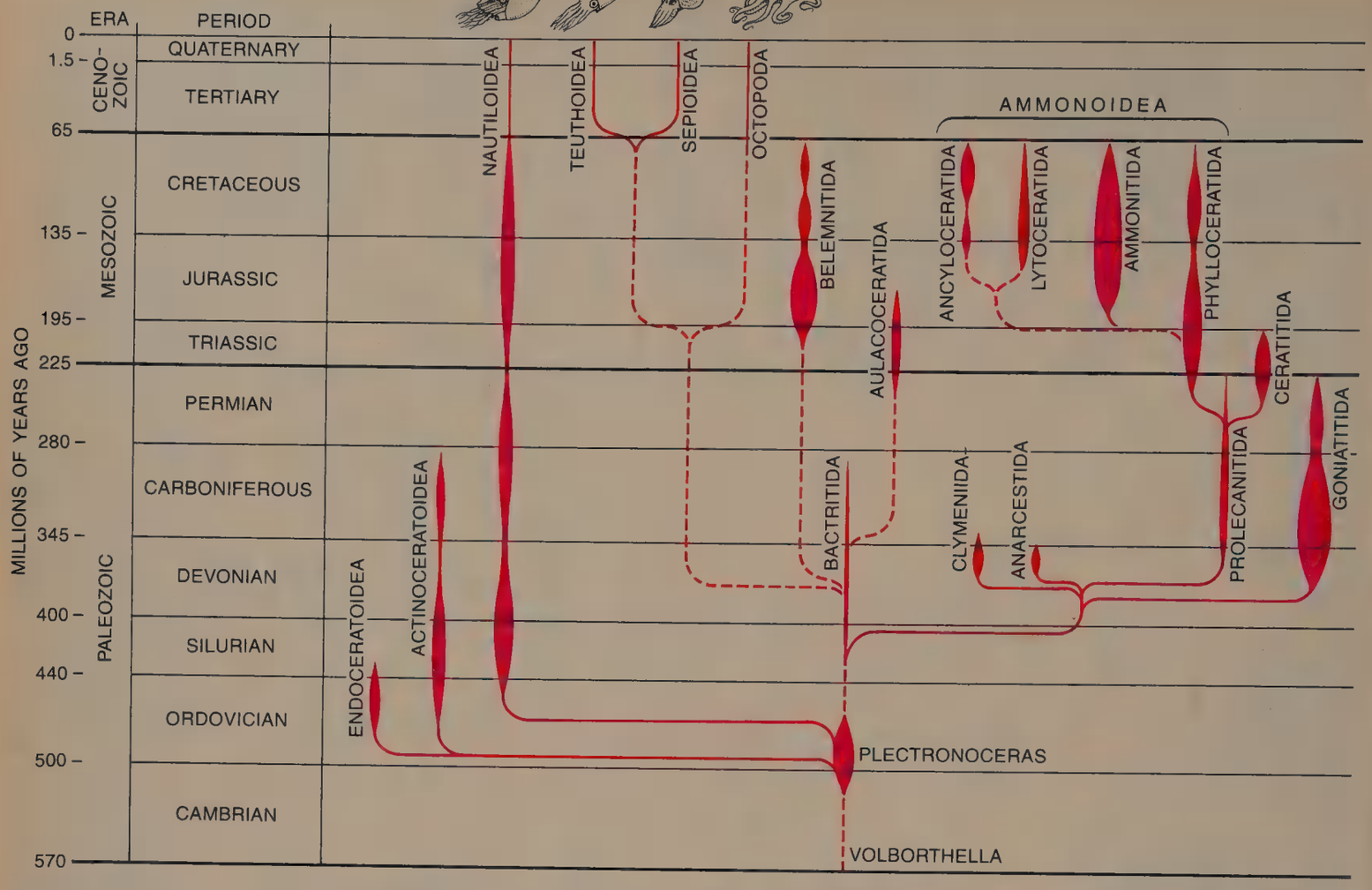
The long evolutionary history of the ammonoids is characterized by explosive radiations and the sudden appearance of many new species, followed quickly by abrupt extinctions. The number of ammonoid species increased rapidly in the Paleozoic, quickly equaling the number of nautiloid species. In the Triassic period of the Mesozoic era, and again in the Jurassic and Cretaceous periods, the ammonoids radiated explosively, eclipsing the nautiloids in number of species. At least three-fourths of the approximately 10,000 known fossil chambered cephalopod species are ammonoids. So diverse are the ammonoids that their fossils are often helpful in dating sedimentary strata.

Perhaps the most striking difference between nautiloid and ammonoid shells is the one between the two organisms' septa (chamber walls) and septal sutures (the line of contact between the septa and the inside of the shell), which are the features of the shell that largely dictate its strength. In the modern nautilus the chambers are emptied of fluid as they are walled off, a contrivance that enables the animal to achieve neutral buoyancy (an overall density or weight per unit volume nearly equal to that of the surrounding seawater) and thus to remain at any depth without effort. The weight of the shell and the body tissues is greater than the weight of seawater, so that the chambers are maintained at

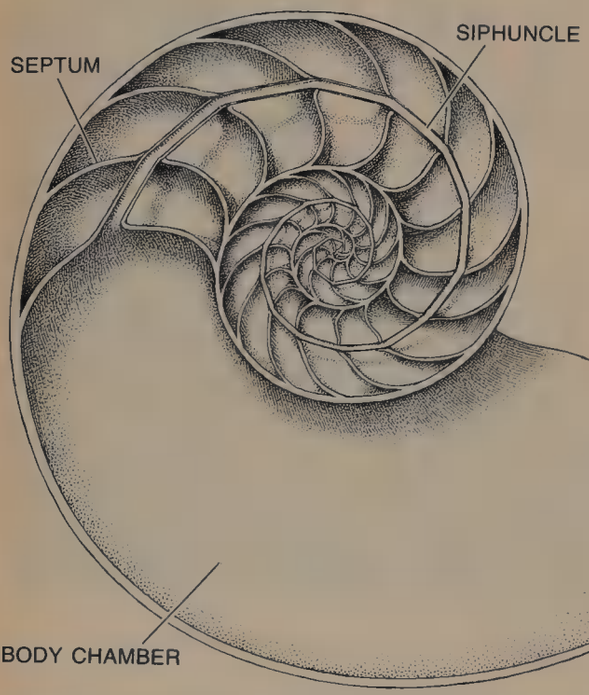


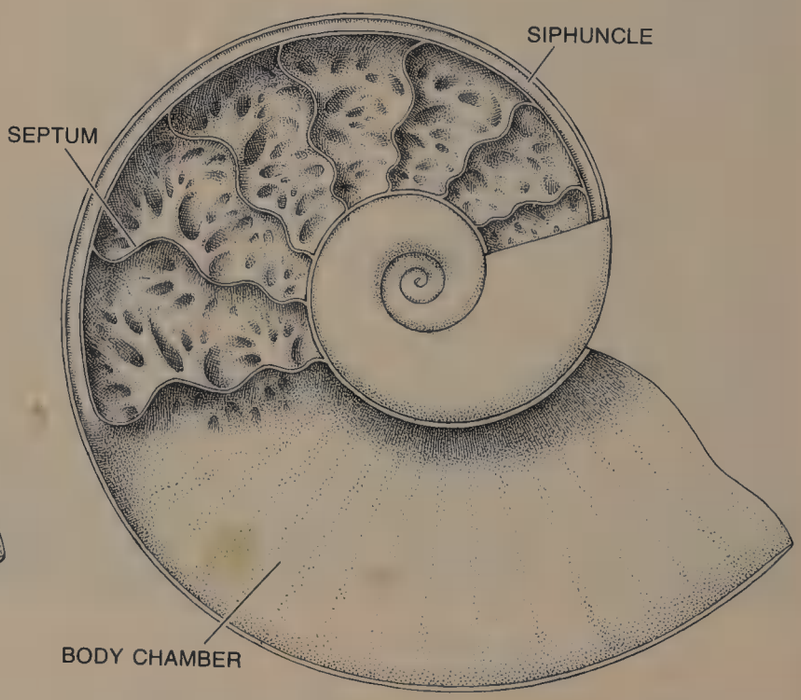
CROSS SECTION OF AN AMMONITE FOSSIL exposes the internal chambers of the shell the animal created by secreting walls behind it as it grew. (The body cavity of the shell is missing in this specimen.) Crystals of chert, a siliceous mineral, deposited on the walls during fossilization give them their layered appearance. This specimen is a member of the genus *Placenticeras,* one of the last genera of ammonites to be extinguished. The form of its shell closely resembles that of the modern nautilus, the only surviving chambered cephalopod. One of the puzzles presented by the extinction of the ammonites is why genera such as *Placenticeras* were extinguished when the ancestors of the nautilus survived. The term ammonite refers only to the last representatives of the group called the ammonoids. The shells of the roughly 7,500 known ammonoid species vary widely in shape and size. The ammonoids were named for the ancient Egyptian god Ammon, often represented as the head of a ram, because some ammonoid shells resemble twisted ribbed ram horns. Most ammonoids are smaller than the specimen shown here, which is 24 centimeters in diameter, but fossils nearly two meters in diameter have been found.





SMALL GROUP OF NAUTILOIDS (the ancestors of the modern chambered nautilus) called the bactritids may be the common ancestor of the extinct ammonoids and belemnites and all the modern cephalopods, including the cuttlefish, the squid and the octopus. The small orthoconic (straight-shelled) bactritids developed into ammonoids when their shells rolled first into loose coils and then into progressively tighter ones. The belemnites had an internal chambered shell surrounded by a soft body. Modern cephalopods have either a chambered internal shell, a laminated cuttlebone, a "pen" or, in the case of the octopus, a vestigial skeleton consisting of small horny supports or a simple rod. These internal hard parts are homologues of the external chambered shells of the nautiloids and the ammonoids.

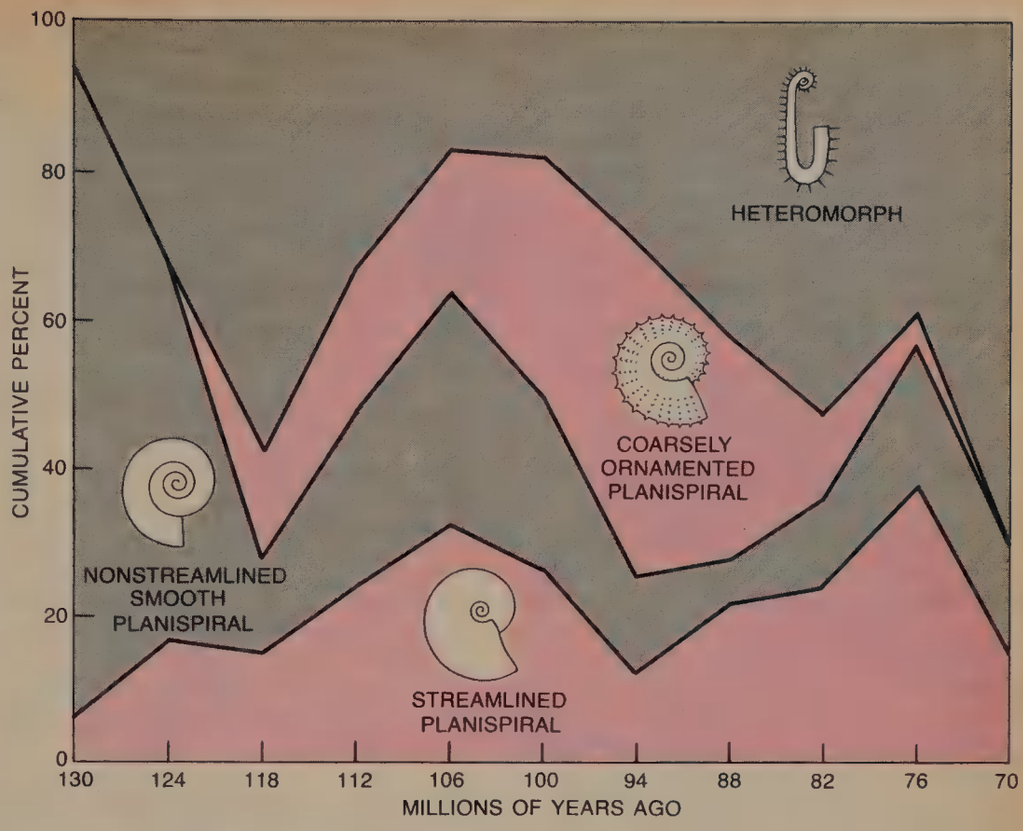




SHELLS OF THE AMMONOIDS differ from those of the nautiloids in two respects. The septa (chamber walls) of the nautiloid shells are smooth, curved surfaces. The septal sutures (lines along which the septa intersect the inner wall of the shell) are gentle curves. The septa of ammonoid shells, on the other hand, are fluted at the periphery, and the septal sutures are folded into complex, frilled curves. The complexity of the septa and septal sutures of the ammonoid shell made it possible for the shell wall to be relatively thin. The other difference between the nautiloids and the ammonoids is the position of the siphuncle, the organ that empties the walled-off chambers of fluid. In the nautiloids the siphuncle generally passes through the center of the whorl; in the ammonoids it is on the outer wall of the whorl.

pressures far lower (less than one atmosphere) than the pressure of seawater (about 40 atmospheres) at the depth where the nautilus is typically found (400 meters). The shells of the nautiloids and the ammonoids were probably also at least partially filled with gas.

The septa of the shells act as braces, giving the shells enough strength to withstand the pressure of seawater at great depths. In most nautiloid shells the septa are smooth surfaces, gently curved in toward the innermost chamber of the shell. The sutures at the intersection of the septa and the shell wall are thickened, but the suture lines are simple curves. In most ammonoids the periphery of the septa is fluted or corrugated. The suture lines are frilled or denticulate (toothed) and of staggering complexity compared with those of nautiloids. Ammonoids and nautiloids are commonly found in the same sedimentary strata, and so they probably lived at similar depths in the ocean. The walls and septa of ammonoid shells are much thinner, however, than those of nautiloid shells of similar diameter. Thus it appears that the architectural complexity of the ammonoid shells made them as strong as nautiloid shells even though they were lighter.



THREE TRENDS IN SHELL TYPE characterize the Cretaceous ammonites. The number of species with streamlined, ornamented or heteromorphic (irregularly coiled) shells increased at the expense of the species with the shell characteristic of most of the earlier ammonoids: a nonstreamlined, smooth, planispiral shell. All three shell types can be interpreted as adaptations to the presence of the more sophisticated marine predators that arose in the Jurassic and Cretaceous periods of the Mesozoic era. Some of the adaptations were more successful than others. Perhaps the most marked (and in some respects the most surprising) trend is the increasing dominance of heteromorph ammonites. The armored ammonites, however, were much less successful; they had almost all been extinguished before the end of the Cretaceous period.

Perhaps as a result the ammonoids produced shell types never evolved by nautiloids, such as shells with a highly compressed or depressed cross section (that is, shells in which either the height or the breadth of the opening in the shell is exaggerated). It is probably also no accident that the variety of shell shapes and the number of ammonoid species increased whenever more complex septal sutures evolved. The very rapidity with which shell shape (and shell ornamentation) changed, however, suggests that both features of the shell were quite plastic and indeed may have had little adaptive significance. Suture lines, on the other hand, are quite conservative, so much so that they serve to establish evolutionary lineages.

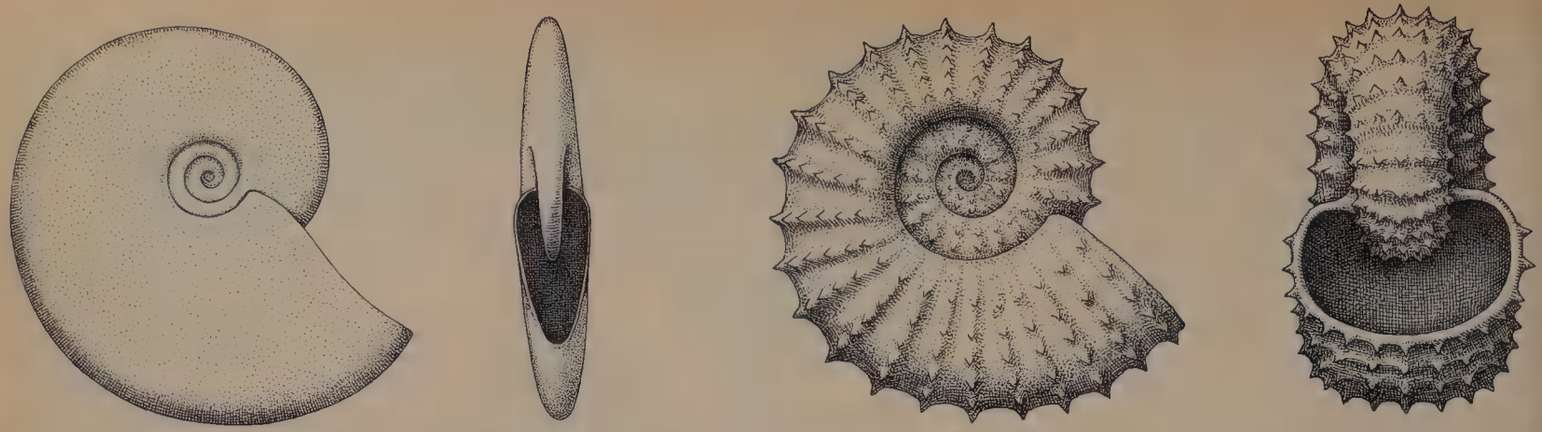
The design of a typical ammonoid shell also suggests that during most of the history of the group shell shape and ornamentation were not dictated solely by their effect on the ability of the animal to elude or discourage predators. All the Paleozoic ammonoids and almost all the ammonoids of the early and middle Mesozoic had planispiral shells, but many of the shells were poorly streamlined and therefore not adapted to fast swimming. Shell ornamentation varied widely, but most shells were smooth or finely ornamented and so gave little protection from predators.

This situation changed markedly in the late Mesozoic, the time of what Geerat J. Vermeij of the University of Maryland at College Park calls the Mesozoic marine revolution. Late Mesozoic radiations of shell-crushing animals such as modern clawed crabs, lobsters, skates and rays, and the arrival of the carnivorous gastropods, bony fishes and many species of large air-breathing marine reptiles, greatly increased the tempo and efficiency of predation in the Jurassic and the Cretaceous. The ammonoids, themselves predators feeding on minute marine animals, small crustaceans and smaller ammonoids, may increasingly have become prey or have lost their prey to more mobile predators. Each of the three trends in ammonite shell type in the Cretaceous is clearly an adaptation to which the selective pressure imposed by the more efficient marine predators had given rise.

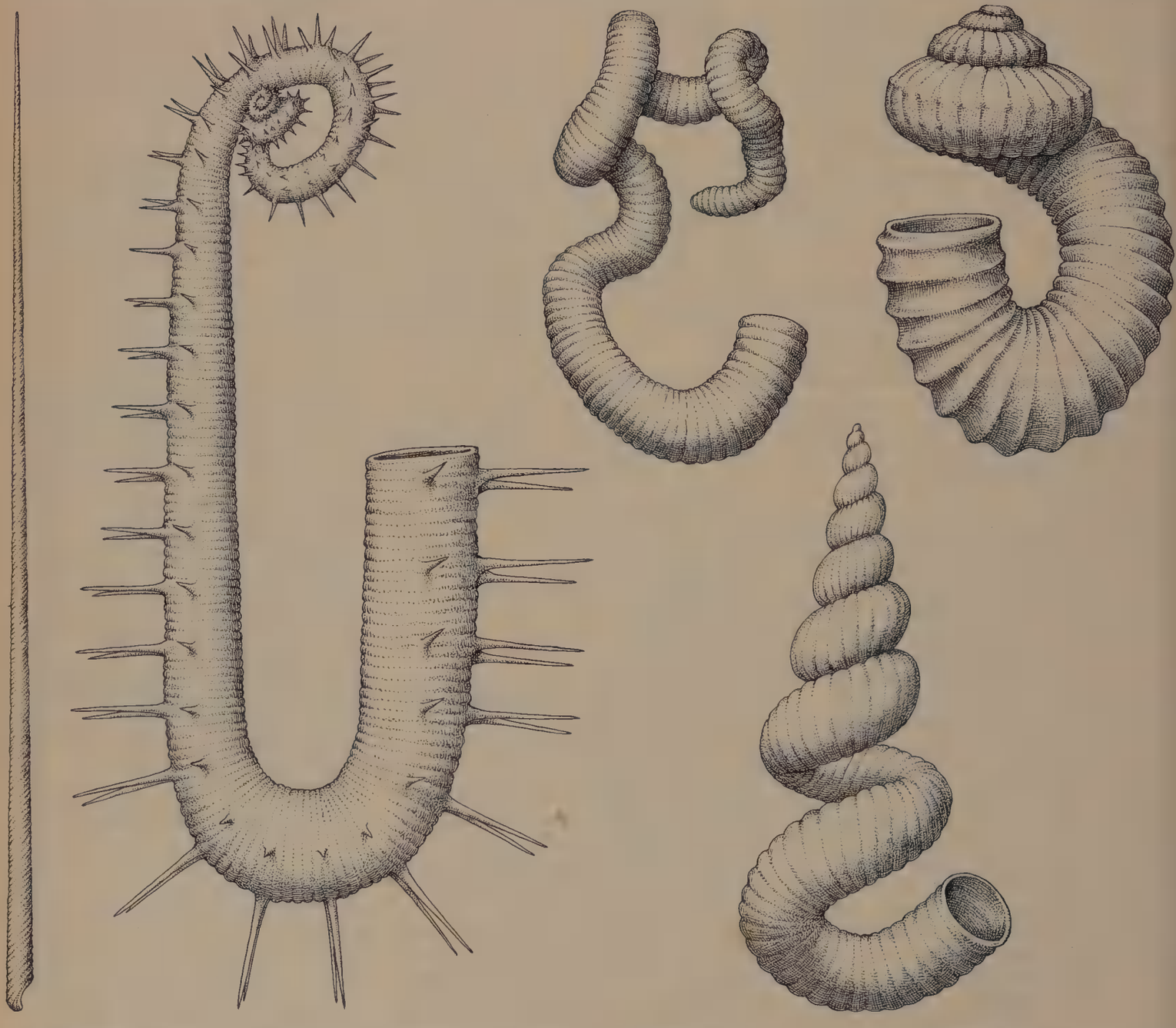
The first trend was toward shell shapes that made the ammonites more agile and faster swimmers. A surprising amount of information about swimming ability can be deduced from shell shape alone. In the 19th century the English mathematician Henry Moseley developed a geometric model for shell coiling based on the logarithmic spiral, a spiral in which the distance from the point of origin increases geometrically with each revolution. Building on this work, David M. Raup of the University of Chicago developed mathematical equations in the 1960's that generate from four parameters of the shell a shape comparable to that of a specimen. When the shell is planispiral rather than helically coiled, only three parameters need be considered: the shape of the closed curve that generates the shell when it is rotated about the coiling axis, often given as the ratio of the breadth to the height of the opening (S), the rate at which the generating curve expands with each revolution (W) and the rate at which the generating curve moves away from the coiling axis with each revolution (D). For example, if the generating curve is a circle, W is large and D is small, the result will be a tightly coiled shell with overlapping whorls.

Raup went on to map ammonoid specimens according to the S, W and D values of their shells. He found that many ammonoid shells were poorly designed for swimming. Two adaptations made the ammonoids (and the nautiloids) free-swimming rather than bottom-crawling: the development of a phragmocone, the evacuated part of the shell that allows the animals to achieve neutral buoyancy, and the development of a hyponome, the funnel-like structure in the body cavity through which water can be forcibly ejected.

One of the problems a swimming ammonoid faces is the problem of attitude control. In an ammonoid the center of buoyancy, which is in the phragmocone, is separated from the center of mass,



STREAMLINED AND HEAVILY ORNAMENTED planispiral shell types are illustrated by these ammonite specimens, a member of the genus *Placenticeras* (*left*) and a member of the genus *Douvilleiceras* (*right*). The smooth exterior and compressed cross section of the *Placenticeras* shell reduced drag, and the tight coiling of the shell increased its hydrodynamic stability; both factors made the animal ■ more agile and faster swimmer. The *Douvilleiceras* shell is less well adapted for swimming because of the rough exterior, the broad cross section, the comparatively loose coiling and the depressed umbilicus (the cavity around the axis of coiling). The spines and thickened ribs of the shell may, however, have served to discourage crabs, marine reptiles such as mosasaurs and other shell-crushing predators.



HETEROMORPH AMMONITE SHELLS are in some instances so bizarre that they were once thought to be the product of "genetic exhaustion," foretelling the extinction of the ammonites. Even now the variety of the shell shapes precludes simple generalizations about their adaptive significance. The heteromorphs shown here are (*from left to right*) members of the genera *Sciponoceras, Anisoceras, Nipponites, Nostoceras* and (*lower right*) *Didymoceras*. Most of the heteromorphs probably led passive, floating lives. Many early Cretaceous heteromorphs seem to have depended on shell ornament for protection against predators. Late Cretaceous heteromorphs, however, particularly those with a *U*-shaped hook at the end of their shell, may have floated at middle depths in the ocean and so escaped the continental-shelf habitats of their bottom-dwelling ancestors and perhaps also the predators that invaded them in the Mesozoic era.

which is in the body cavity. The animals therefore have great static stability. On the other hand, forward thrust from the hyponome is off-axis from the center of buoyancy and tends to rotate the shell away from its equilibrium position; the opposing force generated by the displacement of the mass tends to reorient the shell. The greater the separation between the center of buoyancy and the center of mass, the greater the restoring force and the greater the dynamic stability of the shell. Dynamic stability would be critical to the agility of the animal and would also limit the speed with which it could swim.

A nautilus shell is a good example of a shell in which the centers of buoyancy and mass are widely separated. Raup found, however, that many ammonoid shells, unlike the involute, or tightly coiled, nautilus shell (high *W*, low *D*), are evolute, or openly coiled (low *W*, high *D*). The center of mass and the center of buoyancy tend to be close together in such shells. As a consequence the thrust generated by the squirting of water out of the hyponome is translated into spinning or pinwheeling motion rather than into forward motion.

John A. Chamberlain of Brooklyn College of the City University of New York extended Raup's work on ammonoid shell types, placing particular emphasis on the hydromechanical design of the shells. To hydrodynamic stability he added the further factor of drag: resistance as the shell is pushed or pulled through water. The shell parameter that has the greatest effect on drag is the shape of the generating curve, *S*. If the ratio of the breadth of the shell opening to the height is small, the curve is compressed and the shell is disk-shaped. If the ratio of the breadth of the opening to the height is large, the curve is depressed and the shell is globular. Clearly the smaller the value of *S* is, the less effort is required of the animal to swim at a given velocity.

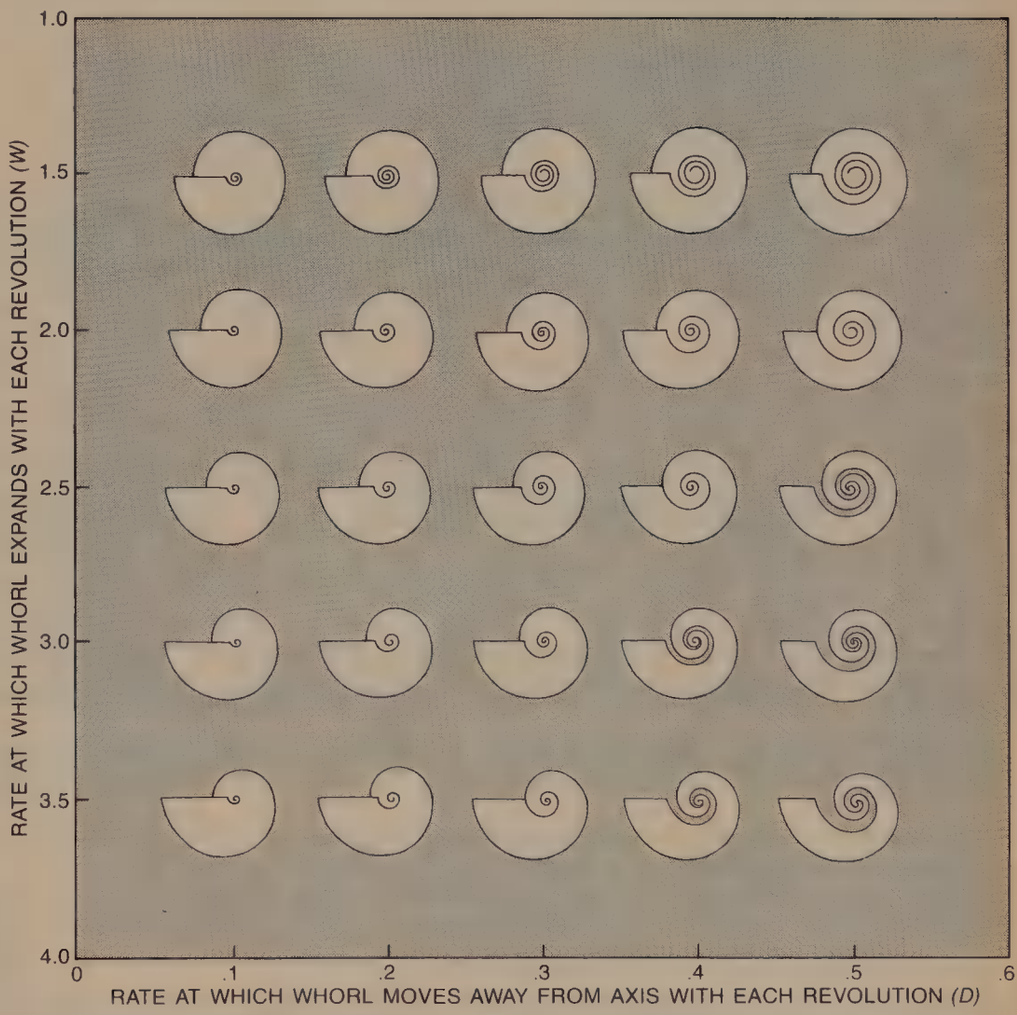
Chamberlain developed equations allowing the swimming velocity of the ammonoids to be calculated from these shell characteristics and others. The shell shape that would enable the animal to reach the greatest velocity would be a compressed shell with overlapping whorls. Chamberlain found that ammonoid shells plotted on *S*, *W* and *D* graphs clustered around the possible shell shapes best designed for swimming, but that great numbers of shells, translating into thousands of species of ammonoids, had shells only minimally or poorly adapted for swimming. These studies suggest that during most of the history of the ammonoids efficient swimming ability was poorly selected for. In the late Cretaceous period, however, highly streamlined shell shapes increased at the expense of less well streamlined ones both regionally and worldwide. Agility and speed had clearly become advantages.

The second trend in ammonite shells is toward increased shell ornamentation among the nonstreamlined species. Over most of their history ammonoids had smooth or finely ornamented shells. Although Chamberlain and Gerd E. G. Westermann of McMaster University have suggested that some shell ornamentation may actually have decreased hydrodynamic drag, just as the pits on golf balls reduce aerodynamic drag, much ammonoid shell ornamentation seems to have no identifiable purpose. Moreover, the pattern of ornamentation is highly variable within a given evolutionary lineage. It would seem, therefore, that barring substantial changes in roughness the pattern of ornamentation on ammonoid shells was unimportant and evolutionarily quite plastic.

In the Cretaceous period, however, those planispiral ammonites that did not have streamlined shells developed large, robust ribs, spines and tubercles. In mollusks living today this type of shell ornamentation appears to have the purpose of defeating attacks by shell-crushing predators. The trend toward heavily armored shells among the ammonites can probably be interpreted in the same way.

How effective was this particular defensive strategy? Many ammonite fossils from the Cretaceous bear the marks of attacks by predators. Some shells have been cut open by crabs. Other fossils bear the tooth marks of mosasaurs and other marine reptiles. In many instances part of the shell near the opening has been broken off. Assuming the animal was not killed, could it survive such an attack? Once again studies of the nautilus provide clues to the answer.

The phragmocone of the nautilus shell serves primarily to buoy up the heavy calcareous shell rather than the



SWIMMING ABILITY OF THE PLANISPIRAL AMMONOIDS can be deduced from the shape of their shells. David M. Raup of the University of Chicago developed equations that will generate any shell shape based on a logarithmic spiral from four parameters of the shell. These computer-generated line drawings show the variation of shell shape with two parameters: the rate at which the cross section of the shell increases with each revolution (*W*) and the rate at which the cross section moves away from the coiling axis with each revolution (*D*). Low *W*, high *D* (evolute) shells are poorly adapted for swimming; high *W*, low *D* (involute) shells are comparatively well adapted. The center of mass of evolute shells, which is in the body cavity of the shell, tends to be close to the center of buoyancy, which is in the phragmocone: the walled-off part of the shell. The two centers are more widely separated in involute shells. Thrust from the directional water jet in the body cavity tends to rotate the shell away from its equilibrium position (in which the center of buoyancy is directly above the center of mass). The greater the separation of the centers of buoyancy and mass is, the faster the shell will reorient itself. Animals with involute planispiral shells therefore tend to be more agile and faster swimmers.



With 3M diskettes, your computer never forgets.

3M diskettes remember everything, every time. Because at 3M, reliability is built into every diskette. We've been in the computer media business for over 30 years. And we've never settled in. We're constantly improving and perfecting our product line, from computer tape and data cartridges to floppy disks.

3M diskettes are made at 3M. That way, we have complete control over the entire manufacturing process. And you can have complete confidence in the reliability of every 3M diskette you buy.

Look in the Yellow Pages under Computer Supplies and Parts for the 3M distributor nearest you. In Canada, write 3M Canada, Inc., London, Ontario. If it's worth remembering, it's worth 3M diskettes.



3M hears you...

3M

body. The density of the shell is much greater than the average tissue density (2.67 grams per cubic centimeter compared with 1.06 grams for the muscles and digestive gland and 1.03 grams for the body fluids). Eric Denton of the Marine Biological Association of the United Kingdom at Plymouth has estimated that about 80 percent of the phragmocone volume is needed just to buoy up the mass of the shell. Thus any loss of shell material causes a comparatively large decrease in the density of the animal. If enough material is removed, the animal is forced to the surface, where it is at the mercy of surface predators and is buffeted by waves and currents.

Many nautilus shells have healed shell breaks, suggesting that the animals have at least some ability to survive sudden losses in weight. The most obvious compensatory strategy would be to refill walled-off chambers with fluid. Until two years ago, however, it was thought that a nautilus at or near the surface had only the option of increasing or decreasing the rate of emptying chambers, not of filling them.

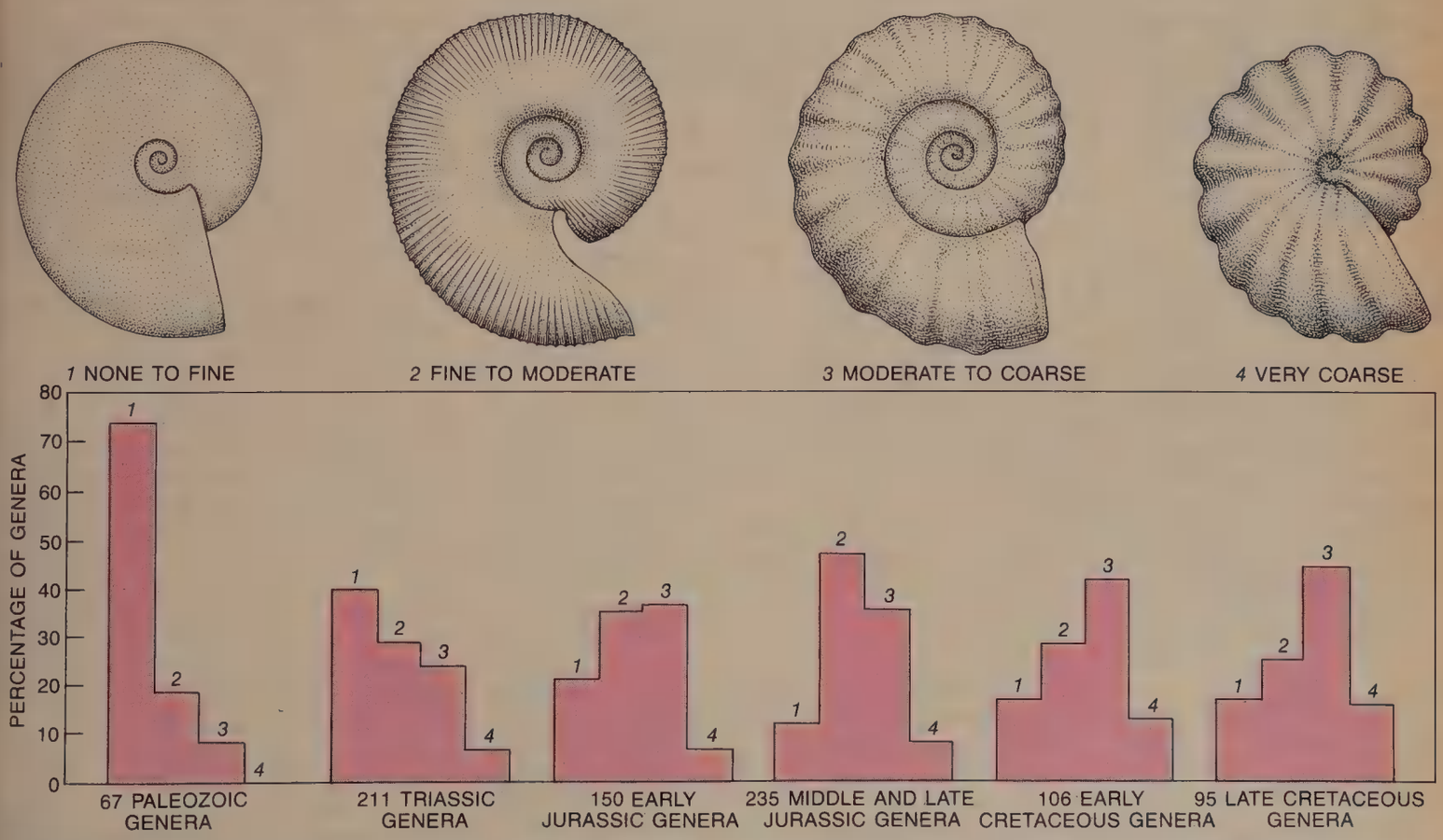
How the nautilus empties walled-off chambers was revealed by Denton and John B. Gilpin-Brown of the Marine Biological Association. They realized in the course of studying modern chambered cephalopods such as the nautilus, the cuttlefish *Sepia* (which has an internal chambered cuttlebone) and the squid *Spirula* (which has a small internal chambered shell) that the chambers in the shell or cuttlebone are originally filled not with gas but with a saline liquid called the cameral fluid. This opened the way to a new interpretation of the siphuncle, the organ that empties the chambers. The mechanism by which the siphuncle empties the chambers was worked out by Denton, Gilpin-Brown, Lewis Greenwald and me.

In the nautilus the siphuncle is a strand of living tissue enclosed in a calcareous tube that spirals through all the chambers of the shell, including the first chamber to arise in the growth of the animal. The substances dissolved in the cameral fluid pass through the porous outer layers of the siphuncle into the cells of the siphuncular epithelium. Enzymes on the folded inner membrane of these cells pump the solute into the fine network of infoldings. When the concentration of ions in the infoldings is sufficiently great, fluid from the chambers is drawn into them by osmosis, draining from there into the blood vessels at the center of the siphuncle.

At the time, never having found a nautilus at or near the surface that was refilling its chambers, we speculated that the part of the siphuncle extending into the first chambers, which are emptied of fluid early, was the vestige of an organ that could both empty and flood the chambers of the ancestral chambered cephalopods. Greenwald and I have since demonstrated that the nautilus itself has this ability. I was recently able to study the phenomenon with freshly captured specimens of *Nautilus macomphalus* in New Caledonia. About one out of 10 nautilus shells in this area shows the scars of large breaks. By simulating such attacks I was able to measure the rate at which the nautilus could refill its shell.

Within the first hour after a nautilus was forced to the surface, liquid slowly began to reenter virtually all the chambers of its shell. The rate of refilling, however, was quite low. The highest rate of weight gain in 25 specimens was less than .5 gram per hour. The total amount of liquid that could be added to the chambers also seemed to be limited. No single chamber was even as much as half refilled and no nautilus (even those having lost as much as 25 grams of shell) ever gained more than five grams in weight. In the nautilus, at least, this adaptation for compensatory buoyancy change seems extremely limited.

What can be said about the ammonites? Many ammonite fossils also show healed shell breaks. In addition the thickened ribs of ammonite shells are oriented radially, so that breaks near the opening of the shell, the shell's most vulnerable part, would be channeled parallel to it and the loss of shell would be minimized. The anatomy and position of the ammonite siphuncle are different from those of the nautilus, but my studies indicate that the rate of chamber emptying and refilling in the nautilus is largely determined by the ratio of the



ROBUST RIBBING, a type of defensive shell ornament, was much commoner in Cretaceous ammonite genera than it was in earlier ammonoid genera. Shell ornamentation is classified as fine or coarse on the basis of the ratio of the width of an outer rib to the shell's diameter.

surface area of the siphuncular epithelium within any given chamber to the volume of the chamber; the ratios are quite comparable in the nautilus and the ammonites. The evidence is therefore inconclusive. Perhaps the ammonites had a refilling system similar in capacity to that of the nautilus; perhaps it was many times more efficient. It is clear, however, that they needed such a system.

The last and most pronounced trend among the Cretaceous ammonites is one toward heteromorphs: species in which the shell forms differ in many ways from the ancestral planispiral form. A few heteromorph species appeared briefly among the thousands of planispiral species in the Triassic and Jurassic periods. Starting about 120 million years ago, however, large numbers of heteromorphs began to evolve. By the end of the Cretaceous period they made up the majority of ammonites, both in numbers and in number of species.

This trend is the most difficult one to interpret. The shell forms of heteromorphs range from openly coiled planispirals to bizarre, seemingly patternless coils. Late in the 19th century and early in the 20th biologists proposed that these aberrant forms were inadaptive end members of the long ammonoid evolutionary history and foretold the final extinction of the ammonites. This notion of "typolysis," or racial senescence, has been discarded; today most workers consider the heteromorphic shell forms to be specialized adaptations to a variety of habitats.

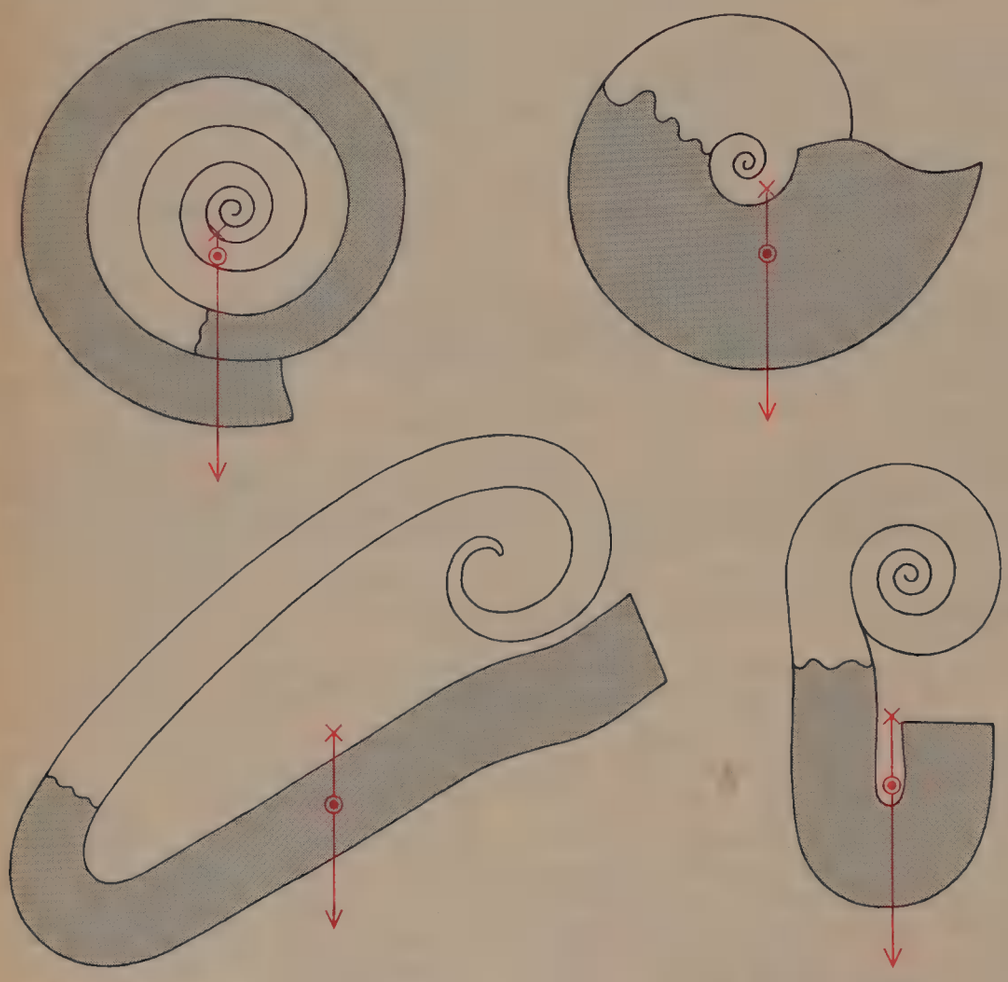
In their diversity of form, however, the heteromorphs defy any simple generalization about the adaptive significance of their shells. Early Cretaceous heteromorph shells ranged from orthocones (straight cones) to openly coiled planispirals to anomalous forms with an extravagant hook at the end of the shell. In the early Cretaceous strata of northern California, perhaps the richest source of early Cretaceous ammonite fossils (soon to be flooded by a dam), many of the heteromorph ammonites are massive and heavily ornamented.

Heavily ornamented planispiral species are found in the same strata. The early Cretaceous heteromorphs, like the heavily ornamented planispiral ammonites, may have arisen from poorly streamlined planispiral species of ammonoids. Jost Wiedmann of the University of Tübingen has shown on the basis of suture lines that heteromorph species in general arose from poorly streamlined planispiral species and often gave rise to planispirally coiled descendants. The plasticity of the shell type suggests that the early heteromorph ammonites, like the ornamented planispiral ammonites, relied for survival primarily on defensive ornamentation rather than on shell form.

By the end of the first half of the Cretaceous many of the larger heteromorph species had disappeared. Although giant forms, such as certain species of the orthoconic genus *Baculites*, reappeared in the late Cretaceous period, late Cretaceous heteromorphs were for the most part smaller and more delicately ornamented than earlier ones. The early Cretaceous heteromorphs, like their ancestors, were probably bottom-dwelling animals, but many of the late Cretaceous species have shell adaptations suggesting that they lived near the surface or at middle depths rather than on the bottom.

Many of the late Cretaceous heteromorph ammonites are helically coiled, and their way of life has traditionally been assumed to be similar to that of the mollusks they most closely resemble: the helically coiled, bottom-dwelling gastropods. The first item of evidence that led to a reinterpretation of their way of life was found by Arthur E. Trueman of the University of Glasgow. In the early 1940's Trueman calculated the centers of mass and buoyancy of a variety of planispiral and heteromorph species of ammonites. His calculations showed that as a rule the heteromorphs had a much greater static stability than the planispirals, largely because the uncoiling of the shell usually separated the center of mass in the body chamber from the center of buoyancy in the phragmocone. Indeed, many of the heteromorphs had such great static stability that they must have been capable of only small changes in orientation.

If the shell of the heteromorph was coiled in such a way that the opening (and thus the body of the animal) faced the bottom, the stability of the shell is not inconsistent with a deep-water existence. An increasing fraction of the heteromorph species that arose in the Cretaceous, however, had a *U*-shaped hook at the end of the body chamber. The phragmocones of these shells were sometimes snail-shaped, sometimes coiled around more than one axis and sometimes nearly patternless in their coiling. In all cases, however, the extreme separation of the phragmocone from the body chamber would have made the shells very stable and would

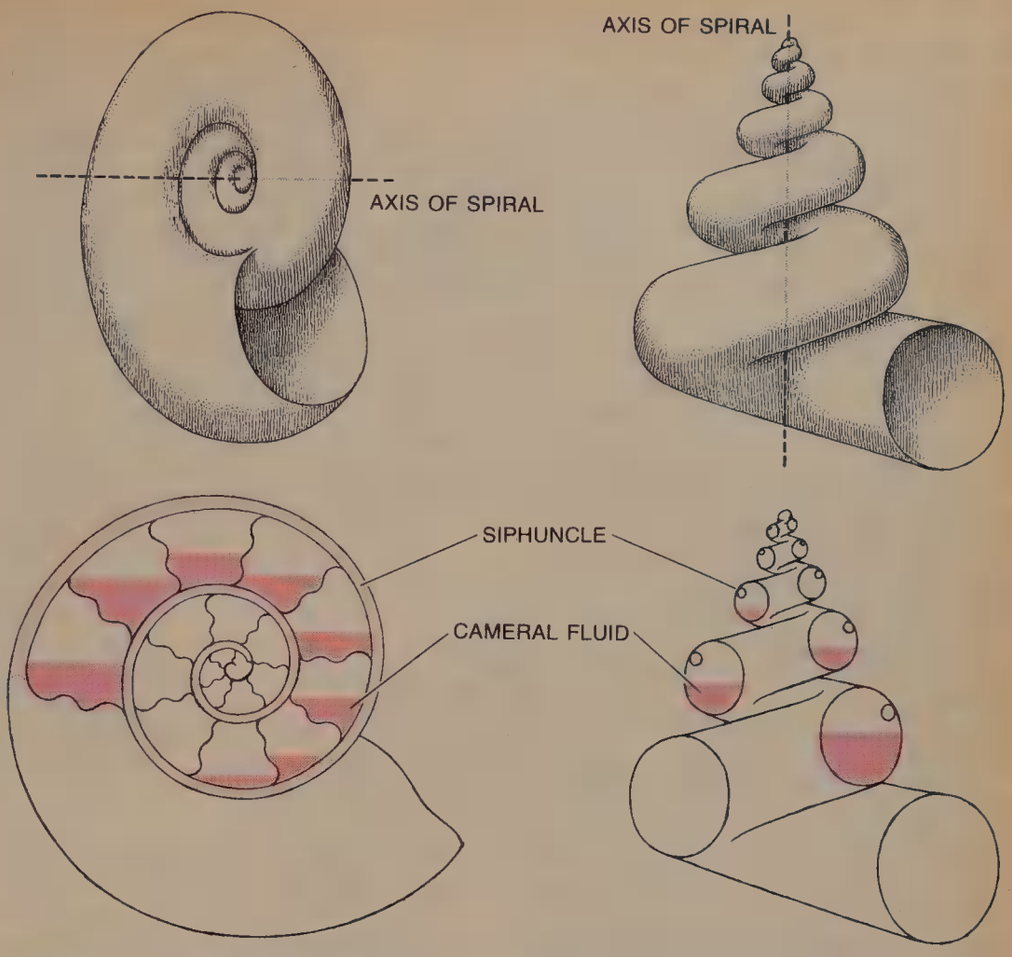


STATIC STABILITY of most heteromorphic shells is even greater than that of involute planispiral shells. Arthur E. Trueman of the University of Glasgow calculated the center of mass and the center of buoyancy for a variety of ammonite shells in the 1940's. The position of the center of mass is dictated by the spatial distribution of the body cavity (*shaded area*) and the position of the center of buoyancy by the spatial distribution of the phragmocone. The static stability of many heteromorphic shells is so great that the animals with these shells were probably capable of only modest changes in orientation. The normal orientation of some of these shells would dictate that the body of the animal face the surface rather than the bottom. This was the first clue that some heteromorphs, unlike their ancestors, were not bottom dwellers.

have dictated that the body of the animal face the surface, an extremely awkward orientation for a bottom dweller.

The next item of evidence toward a reinterpretation of the mode of life of the heteromorphs was provided by Denton and Gilpin-Brown, who pointed out that the siphuncle is positioned in modern cephalopods in such a way that the level of cameral fluid drops below it before the chamber is completely emptied. In the nautilus the siphuncle is in direct contact with the cameral fluid only until the chamber is half empty. After the cameral fluid is "decoupled" (a term introduced by Denton and Gilpin-Brown) the chamber continues to empty because the pellicle, the membrane lining the inside of the chamber, is hydrophilic and acts as a wick to draw fluid up to the siphuncle. The other two modern chambered cephalopod genera, *Sepia* and *Spirula,* also have adaptations for decoupling. It would seem that the chambers could be emptied more efficiently if the siphuncle was oriented so that it was constantly bathed in cameral fluid. Are these adaptations accidental or do they have a functional purpose?

The chambered nautilus lives on the seaward side of reefs where there are often great changes in depth over very short horizontal distances. Bruce Carlson of the Waikiki Aquarium in Hawaii, Mike Weekly of the Seattle Aquarium and I, using remote telemetry to track specimens of *Nautilus belauensis* near the island of Palau in the western Pacific, found that the animals follow the contour of the bottom while foraging, often traveling large vertical distances in a short period. In general the five specimens we tracked moved into shallow water at night and back into deeper water during the day, making daily vertical migrations of as much as 330 meters. *Spirula,* a small squid, also migrates vertically but in mid-water rather than along the bottom.



POSITION OF THE SIPHUNCLE in some heteromorph ammonites may have given them the ability to adjust more easily to changes in depth than planispiral ammonites could, particularly while the animal was growing. Although the anatomy of an ammonite's siphuncle is slightly different from that of a nautilus', the mechanism by which the siphuncle emptied the chambers of fluid was probably substantially the same. Enzymes in the siphuncular epithelial cells establish local osmotic gradients steep enough to draw fluid out of the chambers in spite of the pressure of seawater, which tends to force fluid in the opposite direction. As the pressure changes with depth the osmotic gradient must change if the animal is to continue emptying chambers when it moves to greater depths. If the siphuncle is "decoupled" from fluid in the chambers, however, only the osmotic gradient of the wicklike inner membrane of the shell near the siphuncle need change. In planispiral shells the old chambers rotate up and over as new shell is added at the aperture; the position of the siphuncle in relation to the fluid in any chamber, and thus the degree of decoupling, changes throughout growth. In some heteromorphic shells, such as the torticonic (snail-shaped) shell to the right, the siphuncle in any given chamber remains in a constant position throughout growth and is generally decoupled from the liquid. As a result of this adaptation less energy is required to evacuate chambers.

The process of chamber emptying takes about a month in a small nautilus and three months in larger specimens. It requires energy even if the animal stays at a constant depth, because the hydrostatic pressure of seawater transferred from the body of the animal to the blood in the siphuncle tends to drive fluid back into the chamber, against the osmotic gradient set up by the enzymes in the infoldings of the siphuncular epithelium. As the hydrostatic pressure changes with depth the osmotic gradient must change if the nautilus is to continue emptying a chamber or keep it from being reflooded when the animal moves to a greater depth. It follows that some means of decoupling the cameral fluid from the siphuncle would be an advantage. If when a nautilus changes depth the siphuncle in one of its chambers is decoupled, the salt concentration needs to change only in the fluid absorbed in the chalky tube around the siphuncle and in the wicklike pellicle on the inside of the chamber rather than in the entire volume of cameral fluid.

The decoupling of cameral fluid may have been even more important in ammonites than it is in the nautilus. When a nautilus is mature, virtually all the fluid has been pumped out of its chambers. Recent studies of ammonites, however, suggest that they retained some fluid in their chambers even after they were mature. Richard Reyment of the University of Uppsala has been able to calculate the specific gravity of mature ammonite shells from well-preserved fossils. He found that the shells were much less dense than nautilus shells. Unless the bodies of the ammonites were much denser than the body of the nautilus, 10 to 20 percent of the ammonite phragmocone (compared with 1 or 2 percent of the phragmocone of an adult nautilus) would have had to be filled with fluid to give the animal neutral buoyancy.

The decoupling of cameral fluid in planispiral ammonite species, however, may have been even less efficient than that in the planispiral nautilus. As a planispiral cephalopod adds new shell at the shell opening the older chambers slowly rotate up and over the body chamber. The positions of the older chambers are relatively unimportant in the nautilus both because the chambers are empty and because the siphuncle is centrally located. In the ammonites, however, the older chambers were probably partially filled and the siphuncle was positioned against the outer shell wall. Thus the siphuncle in chambers at the top of the whorl would be decoupled if any fluid had been removed from the chamber, but the siphuncle in chambers at the bottom of the whorl would be

coupled as long as the chamber contained any fluid. The changing position of the siphuncle as the shell grew would have made it even harder for the animal to maintain neutral buoyancy as it changed depth.

One of the major advantages of the heteromorphic shell shapes may have been that the siphuncle in any given chamber remained in a constant position with respect to the cameral fluid as the shell grew. The best examples of this adaptation are the ammonite species belonging to the middle Cretaceous family Turrilitidae and those belonging to the middle to late Cretaceous family Nostoceratidae. In both of these families the shells are helically coiled. In a planispiral shell the axis of coiling is horizontal; in a helicoid shell the axis of coiling is vertical. As a result in the course of shell growth the chambers in a planispiral shell are in effect on a Ferris wheel and the chambers in a helicoid shell are on a merry-go-round.

The siphuncle in a helicoid shell is in a slightly different position with respect to the cameral fluid in successive whorls of the shell but maintains its position as the shell grows; the siphuncle in a planispiral shell does not do so. Moreover, in many species of these two ammonite families the siphuncle migrated high up on the flank of the shell, so that the decoupling of the cameral fluid came soon after the chamber began to be emptied. The migration of the siphuncle, together with the constant position of the siphuncle throughout growth, meant that most of the chambers in the phragmocone were decoupled, even if most of them retained some cameral fluid. Therefore the heteromorphs seem to have gained a more efficient system for the maintenance of neutral buoyancy at the expense of reduced streamlining.

The great stability of the heteromorphic shells and the adaptations they show for the decoupling of cameral fluid have led some workers, such as Andrew Packard of the University of Edinburgh and Herbert Klinger of South Africa, to concur with my supposition that at least some heteromorph species lived at middle depths rather than on the ocean bottom. Many modern cephalopods, such as members of the genus *Spirula* and the cranchid squids, float like balloons at middle depths, moving into shallower waters at night and slowly sinking back into deeper daytime habitats. Some of the heteromorph ammonites may have been capable of such a mode of life and so have escaped from the shallow shelf habitats of the Cretaceous ammonites.

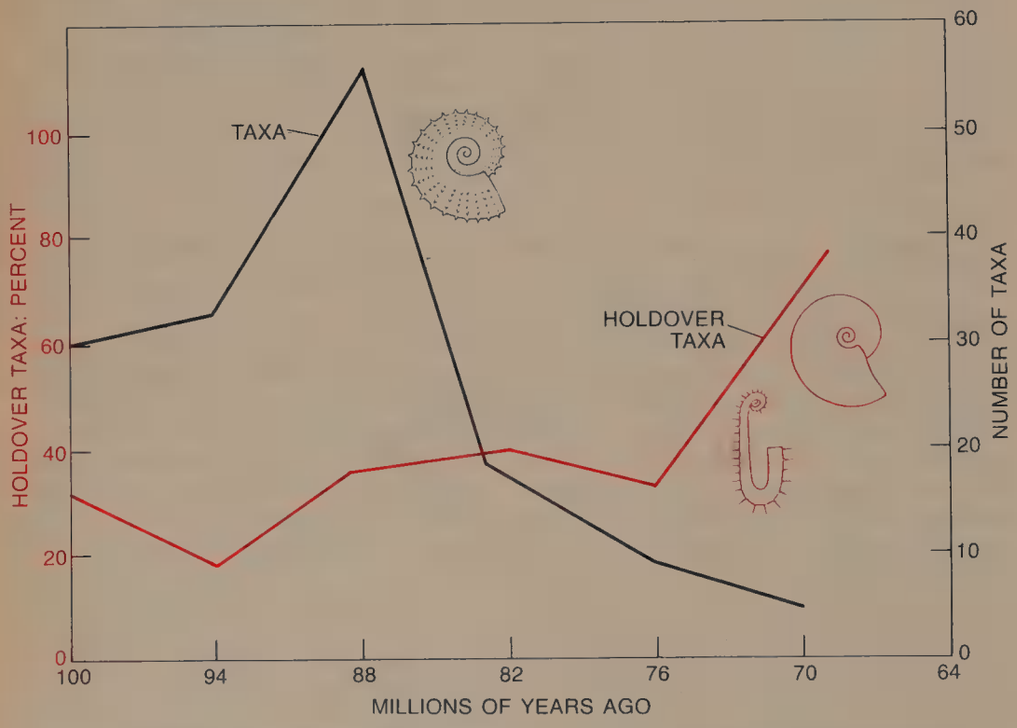
All three shell types (streamlined planispiral shells, heavily ornamented planispiral shells and heteromorphic shells) had appeared at some point in the long history of the ammonoids. What sets the Cretaceous period apart is the sudden increase in the numbers of species with these shell types at the expense of species with poorly streamlined and weakly ornamented planispiral shells, the dominant shell type throughout most of the history of the ammonoids.

In the last stages of the Cretaceous (the Campanian through the Maestrichtian) the number of species of ammonites dropped sharply. There is an interesting pattern to this reduction of diversity. Philip W. Signor of the University of California at Davis and I have analyzed the trends in the origin and extinction rates of families, genera and species of Jurassic and Cretaceous ammonites. The average lifetime for ammonite genera in these periods was about seven million years. William J. Kennedy of the University of Oxford and William A. Cobban of the U.S. Geological Survey have estimated that the average lifetime of ammonite species was between one and two million years. Scattered among these short-lived taxa, however, are a few genera and species that were far more durable; some genera survived for as long as 100 million years. Our studies show that the average lifetime of ammonite genera increased toward the end of the Cretaceous. In other words, the reduction in ammonite diversity was almost entirely due to the extinction of short-lived taxa.

Perhaps not surprisingly, the extinction of short-lived taxa with heavily ornamented and nonstreamlined shells accounts for most of the reduction in diversity at the end of the Cretaceous. That left comparatively few taxa but comparatively long-lived ones, most of which had heteromorphic or streamlined shells. The argument could easily be made that the great reduction in diversity was therefore not necessarily a precursor of the complete extinction of the ammonites.

The fate of the remaining species of ammonites is still not clear. The reason is that the extinction of the ammonites came at a time when shallow ocean basins, the preferred habitat of ammonites, were reduced in area worldwide. Most stratigraphic sections encompassing the transition between the Cretaceous and the Tertiary periods are therefore deep-sea sediments that lack large fossils such as ammonites. So far only two stratigraphic sections that record the final one to two million years of the history of the ammonites have been uncovered. In both of these sections fossils of about 10 species of ammonites have been found in upper Maestrichtian strata. In both roughly half of the fossils belong to genera that have compressed, streamlined shells, such as *Sphenodiscus* and *Pachydiscus*, and half belong to heteromorph genera such as *Scaphites*, *Baculites* and *Diplomoceras*. Unfortunately the stories the sections tell are otherwise contradictory.

At the Stevns Klint section in Denmark, Tove Birkelund of the University of Copenhagen found ammonites of various sizes immediately below strata containing higher than normal concen-



REDUCTION IN DIVERSITY that preceded the complete extinction of the ammonites at the end of the Cretaceous has an interesting pattern. The 300-million-year history of the ammonoids is characterized by the rapid evolution and extinction of many short-lived taxa. The average lifetime of ammonite genera increased during the Cretaceous, however, at the same time that the number of genera dropped precipitously. The reduction in diversity was almost entirely due to the extinction of high-turnover taxa. Why the remaining long-lived (holdover) taxa, some of which closely resembled contemporary nautiloids, became extinct is still unclear.

trations of iridium, one of the platinum-group elements the Alvarezes and their colleagues suggest are derived from a meteorite impact. Recently, however, Wiedmann and I were able to collect large numbers of ammonites from the late Cretaceous section at Zumaya in Spain. As in the Danish section, ammonites are profusely strewn throughout most of the Cretaceous strata; hundreds of fossils can be collected from the beautifully exposed bedding planes.

At about 15 meters below the iridium anomaly in the Zumaya section the number of ammonites gradually falls off. The youngest ammonite in any collection from Zumaya comes from 10 meters below the boundary, even though other large fossils such as echinoids and bivalves are found immediately below and above the boundary. This evidence is negative and could be overturned by the finding of a single new ammonite specimen. Even if a specimen were found immediately below the boundary at Zumaya, however, it would not change the fact the ammonites had virtually disappeared from the region long before the proposed impact of a meteoritic body.

Any hypothesis about the fate of the ammonites must take into account one final question. Why should even the long-lived ammonite species have died out at the end of the Cretaceous when the nautiloids survived? Perhaps no definitive answer will ever be given. My best guess is that the reproductive strategy or perhaps some aspect of the ecology of the adults saved the nautiloids. Plankton, the small floating plants and animals of the sea, had the highest rate of extinction of any group of marine organisms at the end of the Cretaceous; 90 percent of all plankton species were extinguished. Juvenile ammonites hatched from small eggs, with shells no larger than one millimeter in diameter, may have spent their first days or weeks as members of the plankton. The juveniles of nautiloids from the Mesozoic onward seem, on the basis of the form of their shells, to have hatched at a much larger size (from five to 25 millimeters). Juvenile nautiloids probably spent no time as members of the plankton and immediately assumed the near-bottom deep-water foraging mode of life characteristic of the adults. The ammonites may thus have been caught up in the collapse of the plankton ecosystems, either as juveniles or as adults feeding lower down on the food chain than the nautiloids.

For whatever reason, the nautiloids survived the crisis at the end of the Cretaceous and even went on to radiate modestly in the Tertiary period of the Cenozoic era. The low diversity of modern nautiloids suggests, however, that the time for chambered shells, no matter how well engineered, is past.



In by 9:00.



Out by 9:01.

Push one button on the Vivitar® Instant Slide Printer and you'll cut 7,199 minutes off your processing time.

So instead of waiting the normal five days for a photo lab, you'll wait one minute for professional quality color prints. Or half a minute for black and white prints with negatives.

Since the exposure is fully automatic, there are no tricky adjustments to make. And since the printer uses Polaroid® Professional ER instant film, the results will be beautiful.

So why sit there with a handful of slides? In the time it took to read this, you could have already made a print.



The Instant Slide Printer™ by Vivitar

© Vivitar Corp. 1983. In Canada: Vivitar Canada Ltd./Ltée.



RENAULT ALLIANCE WATCHWORD: TECHNOLOGY

If you want to invest in European technology but haven't found a car priced within reach—look at Renault Alliance. Independent suspension with unique twin coaxial torsion bars for a quiet, smooth ride. Sedan comfort for five, with Renault's exclusive pedestal seats for extra rear seat leg room. Power front disc brakes. Electronic fuel injection for an outstanding **52** EST HWY, **38** EPA EST MPG.* Renault Alliance. *Built in America* and priced to bring affordability to European Technology: **$5959.**** It's the one to watch.



Alliance 2-door

* Mileage may vary with speed, trip length, weather. Actual highway and CA figures lower. ** Manufacturer's suggested retail price for 2-door. Tax, license, destination charges, optional or regional equipment extra.

Built by American Motors.

RENAULT

THE ONE TO WATCH

Bilateral Negotiations and the Arms Race

A historical review of the negotiating positions of the U.S. and the U.S.S.R. in several major arms-control talks suggests that each side has special problems in dealing with the other

by Herbert F. York

Why is it so hard for the U.S. and the U.S.S.R. to negotiate mutually beneficial arms-control agreements? The question arises most pointedly in the context of the prolonged bilateral talks now under way in Geneva on the limitation of intermediate-range nuclear weapons in Europe and on the reduction of strategic, or long-range, nuclear arms in general. It applies equally, however, to other attempts—past, present and future—by the two superpowers to resort to direct diplomatic means to moderate the arms race and help reduce the chances of nuclear war. In my view there are certain important idiosyncrasies in the negotiating positions of both parties that make it particularly difficult for them to reach such agreements. To show what I mean I shall review here the historical background and current status of several major lines of arms-control negotiations. In so doing I shall draw in part on my experience as a participant in such sessions, particularly in the latest round of the longest-running nuclear-arms-control discussions of all: the sporadic, 25-year-old effort to achieve a comprehensive ban on nuclear-weapons tests.

The first concrete step toward controlling the nuclear-arms race was the moratorium on nuclear testing that was observed by both the U.S. and the U.S.S.R. from 1958 to 1961. Two main factors lay behind this achievement. One was external: public concern expressed in increasingly urgent terms, both in the U.S. and abroad, about the dangers of radioactive fallout from aboveground nuclear tests, aroused largely by the accidental exposures to radioactivity resulting from the U.S. Bravo test in the Pacific in 1954. The other factor was internal: a deepening concern on the part of President Eisenhower and some of his advisers (and apparently also on the part of Premier Khrushchev and some of his advisers) about where the nuclear-arms race was heading, joined with a determination on both sides to find and take a suitable first step in the direction of doing something about it. The moratorium was not the result of a bilateral negotiation; it was supported by nothing more than a pair of matched unilateral public statements to the effect that "we shall refrain from further nuclear-weapons tests if you will."

In President Eisenhower's view the main purpose for declaring a moratorium was to create a political climate conducive to the negotiation of a detailed treaty on the subject of testing. Unfortunately the negotiations became deadlocked at an early stage over an issue that has continued to plague attempts to negotiate such bilateral arms-control agreements ever since. In brief, the American negotiators were not satisfied with the verification procedures the Russian negotiators were willing to discuss, and the Russians in turn charged that the Americans were actually interested only in spying and wanted far too much in the way of intrusions on their national sovereignty. In this case the issue revolved around the problem of detecting and identifying underground test explosions. The U.S. held that the problem could be handled adequately only on the basis of a system of fairly intrusive and mandatory on-site inspections, and the U.S.S.R. was unwilling to accept any such arrangement.

After negotiations on a treaty had been under way for more than a year President Eisenhower became dissatisfied with their pace, and partly in response to pressure from critics within his administration he announced that the U.S. was no longer bound by its pledge not to test but would not resume testing without giving notice. Several days later Premier Khrushchev declared that in that case the U.S.S.R. was also no longer bound by its pledge not to test but would not begin testing again unless the Western nations did so first. A few months later, in early 1960, France conducted its first nuclear test. Even so, neither the U.S. nor the U.S.S.R. resumed testing immediately. Finally, more than a year later, the U.S.S.R. suddenly initiated a major series of nuclear tests, thus ending the bilateral moratorium.

This action on the part of the U.S.S.R. has often been cited as an example of Communist perfidy. It was indeed wrong, I believe, but it was not perfidious, since no agreement—either tacit or explicit—to refrain from such tests existed at the time. In any case the U.S. responded quickly with a major test series of its own. Finally, as a result of the alarm produced by the Cuban missile crisis of October, 1962, negotiations were resumed, and in 1963 President Kennedy and Premier Khrushchev finessed the problem of how to verify a ban on underground tests by negotiating and signing the Limited Test Ban Treaty, which banned tests in the atmosphere, outer space and underwater but allowed them to continue underground.

Following this qualified success the U.S. and the U.S.S.R. continued to explore the possibility of a ban on all nuclear tests, but they were unable to resolve the fundamental issue of verification. Some additional partial measures were eventually achieved, however, including the Threshold Test Ban Treaty of 1974 and the Peaceful Nuclear Explosions Treaty of 1976. In addition the closely related and very important Non-Proliferation Treaty and the treaty establishing a Latin-American nuclear-weapon-free zone (the Treaty of Tlatelolco) were signed, both in 1968. The goal of a comprehensive test ban, however, has remained elusive.

In 1977 President Carter included the negotiation of a comprehensive test ban among his top arms-control priorities. By that time the position of the U.S.S.R. with regard to on-site inspections had evolved to the point where it was willing to accept a voluntary form of on-site

inspection; meanwhile, the U.S. position had evolved to the point where it was willing to accept a carefully hedged form of voluntary on-site inspections instead of mandatory ones. In addition the Russians indicated their willingness to accept a substantial number of specially designed and constructed "national seismic stations" on their territory. Between 10 and 15 of these stations were to be built according to agreed specifications and provided with cryptologic systems that would guarantee that the data stream received from them was continuous and unmodified.

The details of a treaty were only about half worked out when external events slowed the negotiations to such an extent that it became impossible to complete the process before the end of President Carter's term. Among the external factors were unanticipated difficulties in the second phase of the strategic-arms-limitation talks (SALT II), which were going on at the time, the seizure of U.S. hostages in Iran and the Russian intervention in Afghanistan. The Carter round of test-ban negotiations was adjourned indefinitely a week after the 1980 elections.

The Reagan Administration decided immediately after taking office not to resume the negotiations, but it continued to debate for almost a year and a half its reasons for not doing so. One group argued that the administration should simply declare that as long as the U.S. relied on nuclear weapons as an important element of its defense strategy, it would be necessary to continue testing them, and so a comprehensive test ban would not be in the nation's interest for the foreseeable future. A second group argued that the main problem was that there still was no adequate system for verifying a ban on underground tests (the implication being that if there were, the U.S. might then be willing to negotiate an agreement). The second group eventually won the internal debate, and the current official position is that a comprehensive test ban remains a "long-term goal" of the U.S. but that "international conditions are not now propitious for immediate action on this worthy project."

In recent years, including the entire tenure of the Carter Administration, the Joint Chiefs of Staff have consistently and forcefully argued that a comprehensive test ban would not be in the best interests of the U.S., whether or not it could be adequately verified (and they did not believe it could be). In particular, they asserted that as long as the nation maintains a stockpile of nuclear weapons it will continue to be necessary to conduct at least occasional tests to be confident that the weapons in the stockpile are still in working order. The Joint Chiefs based this position on the advice given to them by their own advisers in the Defense Nuclear Agency and on the advice of most of the experts in the nuclear-weapons laboratories. The issue of "stockpile reliability," as it is called, has continued to be the principal basis for the opposition to a comprehensive test ban in both the Department of Defense and the Department of Energy.

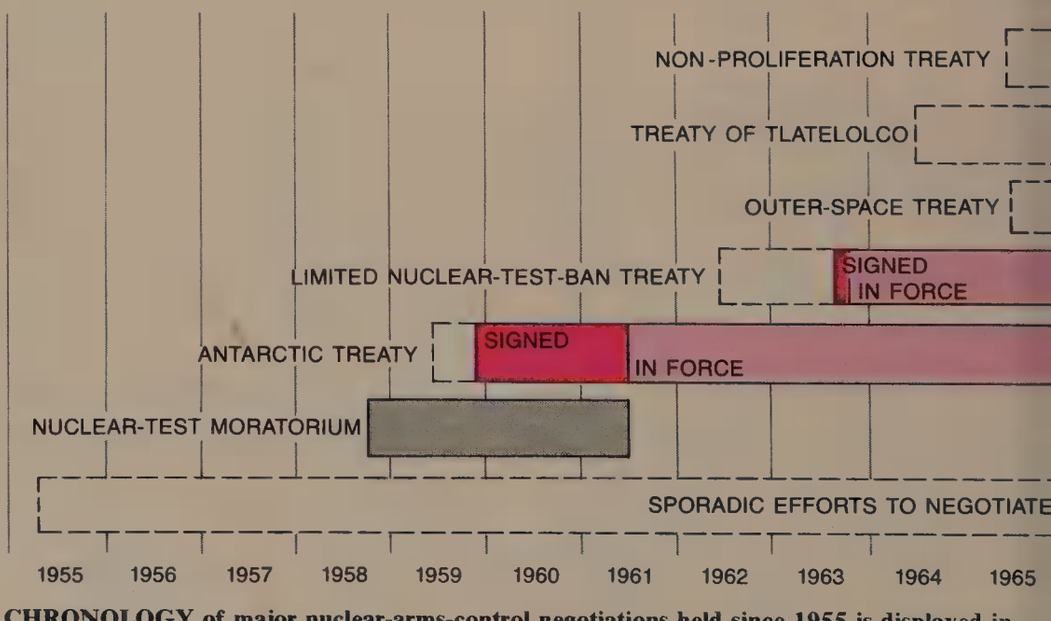
The main argument in favor of a comprehensive test ban in recent years has been that it is an essential element of the nation's nonproliferation policy. In particular, it is usually pointed out that the Non-Proliferation Treaty, which has been signed or acceded to by a large majority of the world's nations, calls for "good faith" negotiations by the superpowers to end the arms race and ultimately to eliminate their nuclear weapons, and the seriousness with which they approach the test-ban issue is widely taken as a measure of their good faith in the matter.

The verification issue remains highly controversial. In brief, large- and moderate-yield nuclear explosions can be readily detected and identified by means of remote sensors but very small ones cannot. The boundary separating the two classes of events is indistinct, however, and that leads to widely varying interpretations of the data, depending on the predisposition of the person interpreting them.

From the beginning of the nuclear age the dominant view in both the U.S. and the U.S.S.R. has been that the interests of both countries and indeed of the world would be best served if there were no other nuclear powers, or at any rate as few as possible. This is one of the rare instances in which both the Americans and the Russians seem to be fully aware that their interests are exactly parallel.

The U.S. has devised a number of specific policies and actions designed to promote the goal of the nonproliferation of nuclear weapons, sometimes unilaterally and sometimes in concert with other nations. Among the measures taken were the creation of the International Atomic Energy Agency, the passage by Congress of the Nonproliferation Act of 1978 and numerous diplomatic initiatives involving such matters as severely limiting the reprocessing of spent fuel from nuclear reactors and placing restrictions on the sale of nuclear-power equipment. Last but certainly not least there is the Non-Proliferation Treaty of 1968, which went into force in 1970.

In essence the Non-Proliferation Treaty can be viewed as an attempt to divide the nations of the world permanently into two categories: one consisting of those nations that already had nuclear weapons and that were pledged not to help anyone else obtain them, and the other made up of those nations that did not have nuclear weapons then and promised to forgo them forever. In return for the promise not to acquire nuclear weapons the non-nuclear nations exacted two promises from those signatories that already had them. One promise was that the "haves" would help the "have-nots" acquire the technology needed to enable them to benefit fully from the peaceful applications of nuclear energy; the other was that the "haves" would undertake serious negotiations to



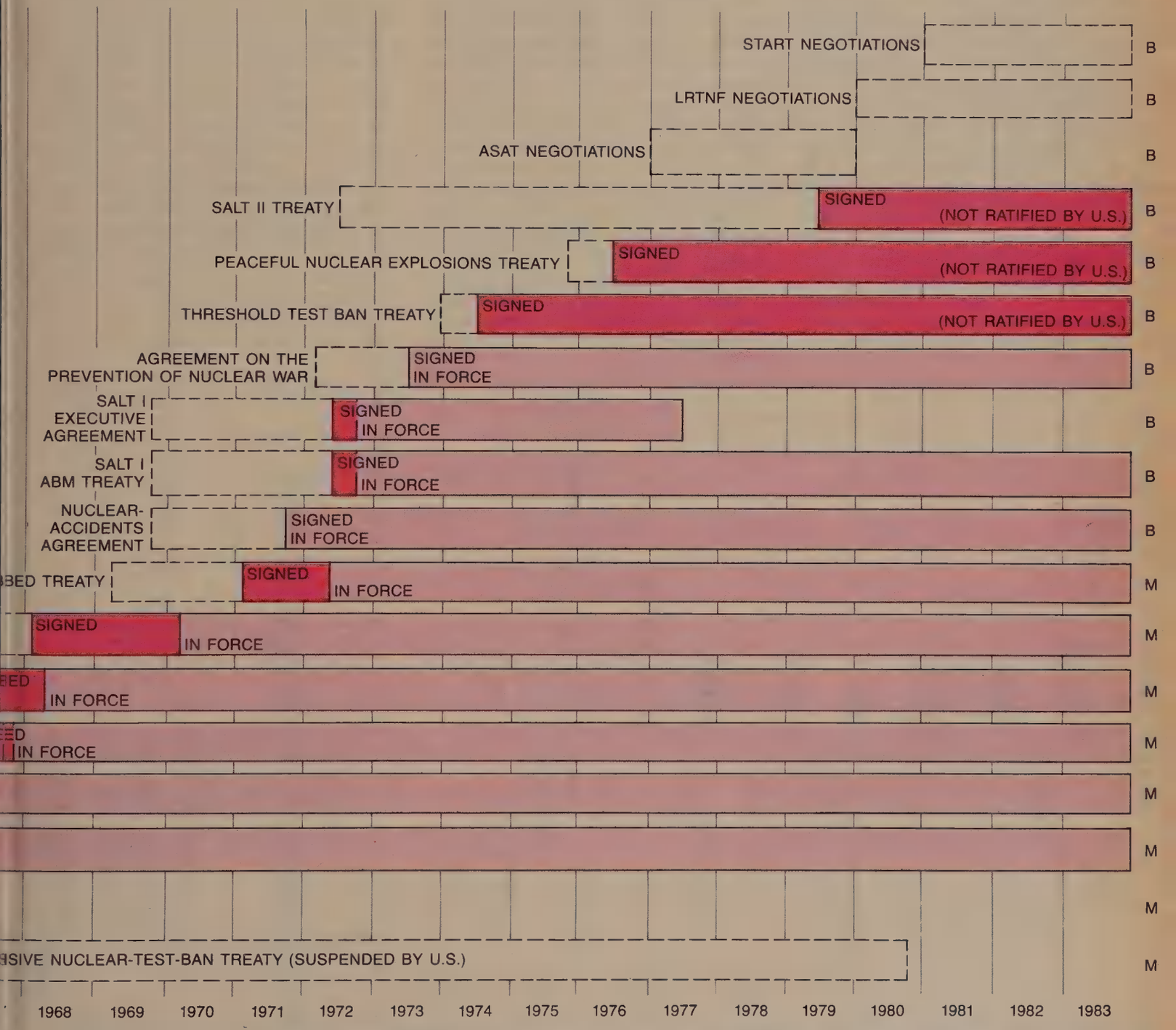
CHRONOLOGY of major nuclear-arms-control negotiations held since 1955 is displayed in the chart on these two pages. The part of each bar represented by the broken outline shows the duration of the negotiating period, the dark-colored part corresponds to the interval between the signing of a treaty and its ratification and the light-colored part is the period during which an agreement is in force. The capital letters at the extreme right indicate whether an agreement is multilateral (*M*) or bilateral (*B*) between the U.S. and the U.S.S.R. Three bilateral treaties

end the nuclear-arms race and eventually eliminate their own nuclear weapons. The treaty also calls for a review of the overall situation every five years. At both of the first two review conferences (in 1975 and 1980) many participating countries complained that neither of the superpowers was living up to the special obligations described above. In spite of these complaints there have been no defections from the treaty, and more importantly there has been no further nuclear-weapons proliferation since the treaty was signed.

There are of course cases where further proliferation could happen soon (India, Pakistan, Israel, South Africa, Argentina and Brazil, for example), but so far even those countries that have evidently advanced quite far down the road toward a nuclear-weapons capability have refrained from testing or otherwise overtly establishing themselves as nuclear-weapons states. (This observation applies even to India, which exploded a nuclear device 10 years ago but has not created a nuclear-weapons force.) Surely the policies and actions described here, including in particular the Non-Proliferation Treaty, must be viewed as a large part of the reason for the long and quite unexpected delay in the appearance of additional nuclear powers beyond the five that have had a nuclear-weapons force since China became a member of the club in 1964.

In spite of this generally successful record it must be noted that some very important nations have refused to sign the Non-Proliferation Treaty, including China, France, Cuba, India, Argentina, Brazil and Israel. In the case of France and China the omission is particularly serious, but at least each of them has pledged to live up to the spirit of the treaty even while refusing for political reasons to formally adhere to it.

President Johnson early in his term of



remain unratified by the U.S. Senate: the Threshold Test Ban Treaty (signed by President Nixon in 1974), the Peaceful Nuclear Explosions Treaty (signed by President Ford in 1976) and the SALT II Treaty (signed by President Carter in 1979). The U.S. and the U.S.S.R. are currently engaged in two nuclear-arms-control negotiations in Geneva: the Strategic Arms Reduction Talks (START) and discussions on the limitation of intermediate-range nuclear weapons in Europe. (Such weapons are technically known as long-range theater nuclear forces, or LRTNF.) Negotiations aimed at achieving a comprehensive nuclear-test ban and an agreement to forestall the further development and deployment of antisatellite (ASAT) weapons have apparently been abandoned by the Reagan Administration.

office proposed a freeze on the further development or deployment of strategic delivery systems. After some discussion of the proposal at the Geneva arms talks, President Johnson met with Premier Kosygin in Glassboro, N.J., where they discussed the matter privately. It was on this occasion that Secretary of Defense Robert S. McNamara, who was also present, first formally raised the idea of initiating such a freeze by prohibiting the deployment of anti-ballistic-missile (ABM) systems. Premier Kosygin rejected the idea on the ground that ABM weapons were purely defensive and that it was, after all, only offensive weapons that threatened the lives of many millions of people. Secretary McNamara and other Americans, however, persisted in arguing at various public and private meetings that the development and deployment of ABM systems stimulated the arms race just as much as the development and deployment of offensive systems because of an "action-reaction cycle" in which the development of defensive weapons promoted the development of new offensive weapons. Eventually the Russians came around to this point of view.

President Johnson's attempts to initiate formal bilateral negotiations on the limitation of strategic arms were thwarted by the Russian intervention in Czechoslovakia in 1968. Only after President Nixon took office in 1969 and the international scene had calmed down was it possible to get what is now known as the SALT process under way. The first series of these talks (SALT I), concluded in 1972, resulted in two achievements: a treaty that severely limited the deployment of ABM systems and an executive agreement that temporarily froze the deployment of land-based and sea-based offensive missiles at numbers equal to those already deployed plus those whose deployment was under way at the time the agreement was signed. The net result was a rough strategic balance that was intended not to be an end in itself but to serve as the basis for further negotiations whose ultimate aim was to produce a situation of actual overall parity at a much lower level of total deployments.

One of the principal residual problems of the SALT I agreements was that the U.S.S.R. was left with a substantial number of very large missiles (designated SS-18's by Western sources), whereas the U.S. had no missiles of comparable size. In the early 1970's, before highly accurate MIRV's (multiple independently targetable reentry vehicles) were fully developed and widely deployed, this imbalance did not seem serious to most observers, but in recent years, with the perfection of high-multiplicity MIRV systems and the achievement of very high accuracies by the U.S.S.R. as well as by the U.S., the issue of the SS-18's has come to be widely seen as a serious matter. The new developments increase the vulnerability of the land-based component of the U.S. "triad," the leg on which U.S. defense policy places the most reliance.

The next step after SALT I was the Vladivostok Agreement of 1974, signed by President Ford and President Brezhnev. In essence it was intended to establish general guidelines for the detailed SALT II negotiations, which had already started in Geneva. The guidelines called for an "equal aggregate limit on delivery vehicles" of 2,400 and an "equal aggregate limit on MIRV systems" of 1,320. The effort to transform the Vladivostok guidelines promptly into a formal treaty ran afoul of both technical and political problems. The technical problems had to do with cruise missiles and the Russian "Backfire" bomber; the political problems arose from increasing disillusionment about détente and the challenge to President Ford's nomination by the right wing of the Republican party.

In 1977 the SALT II negotiations became the principal element of Presi-



NUCLEAR TESTING by the U.S. (*black bars*), the U.S.S.R. (*colored bars*) and all other nations (*white bars*) is charted from 1951 through 1982. The gap at the left reflects the moratorium on nuclear-weapons tests that lasted from 1958 to 1961. The net effect of the Limited

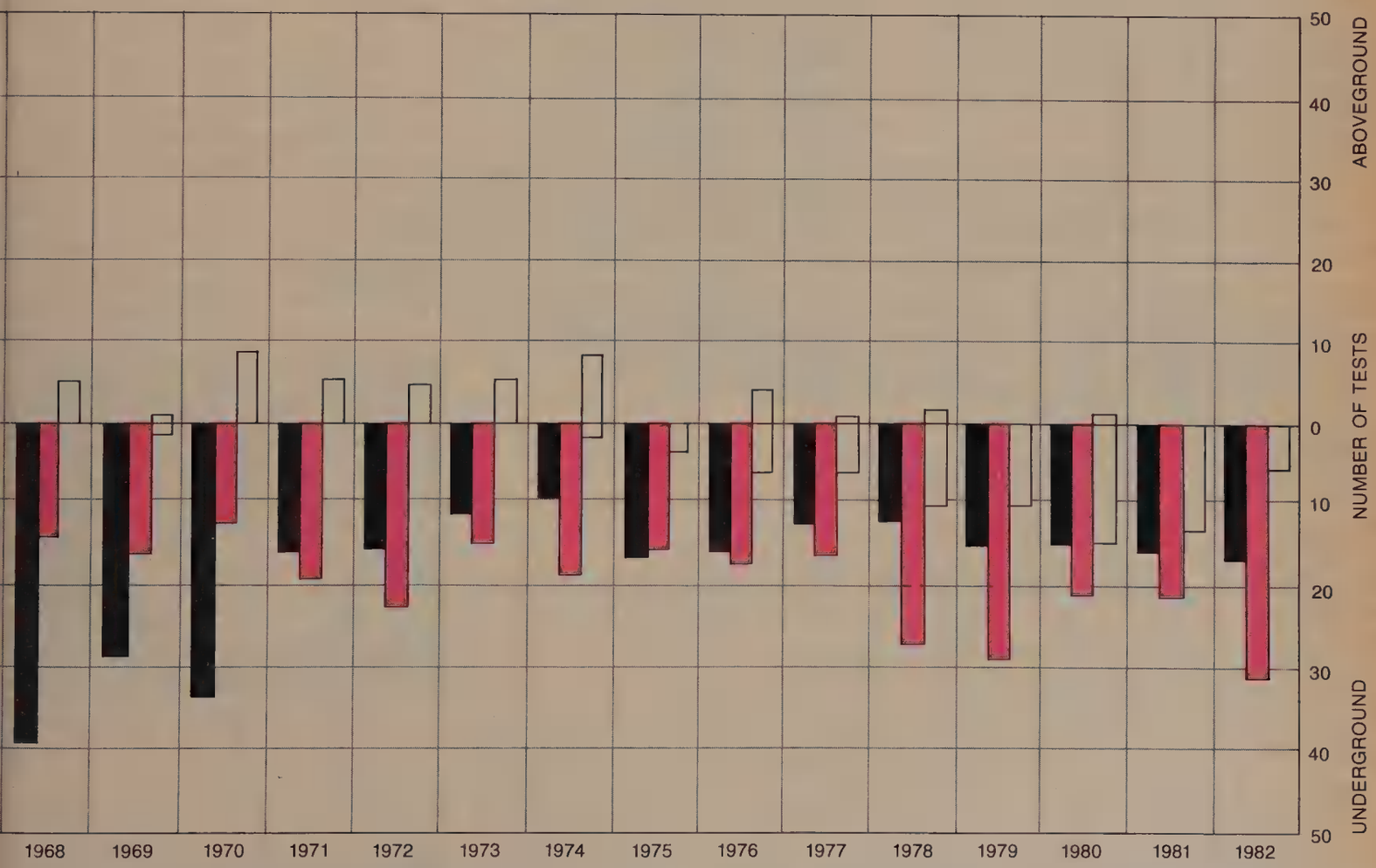
dent Carter's arms-control policy. At first President Carter proposed some reductions in the number of deployed systems below the Vladivostok aggregates, a 50 percent cut in the Russian heavy missiles (modern versions of which were barred to the U.S. entirely) and limits on tests, improvements and numbers of MIRVed intercontinental ballistic missiles (ICBM's). The latter provisions were intended to forestall any further worsening of the problem of ICBM vulnerability and generally reduce technological pressures on the arms race. The U.S.S.R., however, insisted on staying within the Vladivostok framework. The SALT II negotiations were completed in Geneva and in a series of high-level meetings over the next two years, and the resulting treaty was signed by President Carter and President Brezhnev in Vienna in 1979.

The SALT II Treaty basically kept the Vladivostok framework. It set an overall limit on strategic nuclear delivery vehicles of all kinds at 2,250 for each side, and within that limit it provided a nested series of sublimits, setting a ceiling of 1,320 on MIRVed systems of all kinds (including heavy bombers with long-range cruise missiles), a ceiling of 1,200 on MIRVed ballistic missiles and a ceiling of 820 on MIRVed ICBM's. It also limited each side to only one new type of ICBM, banned major changes in existing systems and set specific limits on the number of reentry vehicles that could be placed on the one new ICBM and on each type of existing ICBM. In addition temporary limits on ground-launched cruise missiles (GLCM's), sea-launched cruise missiles (SLCM's) and mobile ICBM's were intended to provide time for the negotiation of longer-term agreements on these systems.

Much of the criticism of the SALT II Treaty in the U.S. grew out of widespread frustration over the nation's international troubles, exaggerated fears of nuclear inferiority (stimulated by right-wing critics at home) and lack of confidence in President Carter's competence in security matters generally and his commitment to military programs in particular. In addition the impending presidential election made Republican senators reluctant to hand a major foreign-policy victory to a weakened opponent. In spite of broadly based satisfaction with the overall terms of the treaty itself, there were several particular treaty issues that raised serious questions in the minds of those senators and other political figures who took an extremely cautious view of the entire arms-control process.

One such issue was that SALT II, like SALT I before it, did nothing to alleviate the heavy-missile problem other than to put a rather high cap on the number of warheads each type of missile could carry. A second issue involved the Backfire bomber, an aircraft that is technically capable of reaching the U.S. under certain unusual circumstances. Some American observers asserted that the Backfire was therefore an intercontinental bomber that must be included under the ceilings. The U.S.S.R. contended (and many American observers agreed) that the Backfire's mission was only that of an intermediate-range bomber, and so it should not be included, particularly since the U.S. medium-range aircraft in Europe capable of reaching the U.S.S.R. had been excluded at Vladivostok over the strong objections of the U.S.S.R. A third issue involved the encryption of certain test data broadcast by Russian missiles during test flights. This last subject is a rather arcane one that cannot be usefully elaborated in an unclassified discussion. Suffice it to say that most of the professionals who were then involved believed the matter had been adequately handled.

On the Russian side the principal issue raised after the Vladivostok meeting concerned cruise missiles, which the U.S.S.R. wanted sharply constrained. (To avoid limitations on exist-



Test Ban Treaty of 1963 was not to diminish the rate of nuclear testing but merely to drive most of it underground. (Only France and China have tested in the atmosphere since then.) The figures were supplied by the Stockholm International Peace Research Institute.

Medicine is always changing, now there's a text that changes right along with it.

SCIENTIFIC AMERICAN *Medicine.*

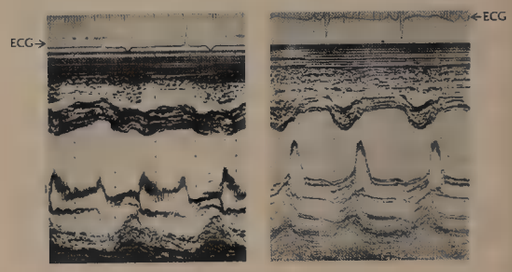
The answers that you get from SCIENTIFIC AMERICAN *Medicine* are current, because each month new chapters replace the old. You will know where to look for the answers, because this advanced text follows the familiar structure of medical knowledge, changing as medicine changes.

Comprehensive. Leading authorities from Harvard and Stanford cover the full range of medicine for you the seasoned clinician.

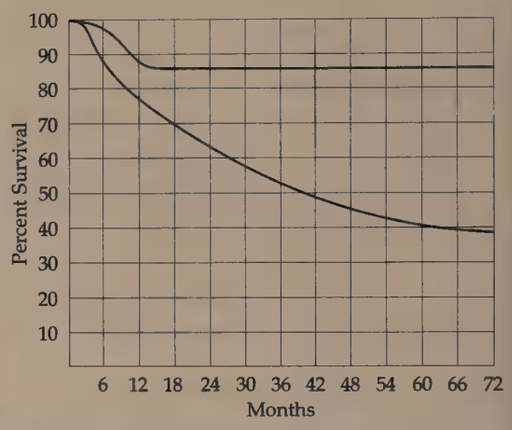
Complimentary CME. At no extra cost, we offer a CME program in which you can earn 32 top credits* per year.

Trial Offer. We invite you to try SCIENTIFIC AMERICAN *Medicine* for two months with our compliments. Send us the coupon and you will receive the two-volume text and two monthly updates. You may also take a CME test for credit. At the end of 60 days, if you decide to continue the subscription, you will owe us $220 for the first 12 months (current renewal is $170); otherwise, return the books; there is no charge.

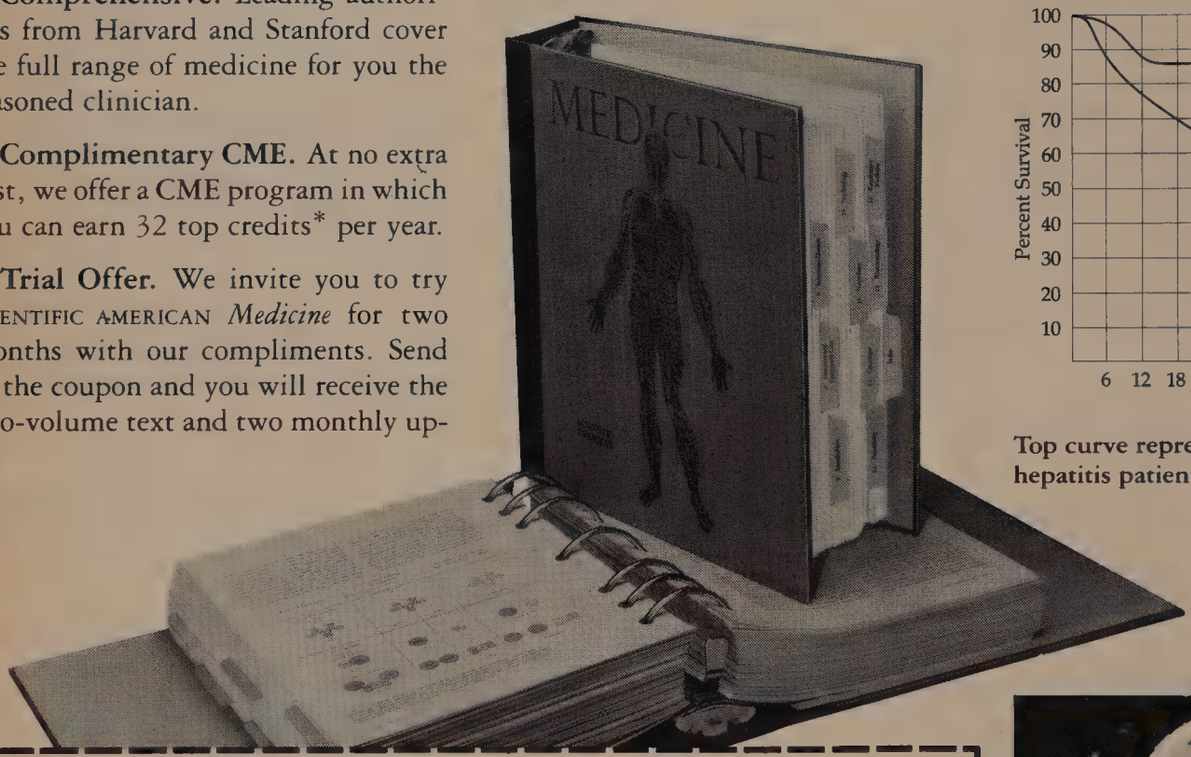
Please mail the coupon today so that we can put medical progress to work for you.



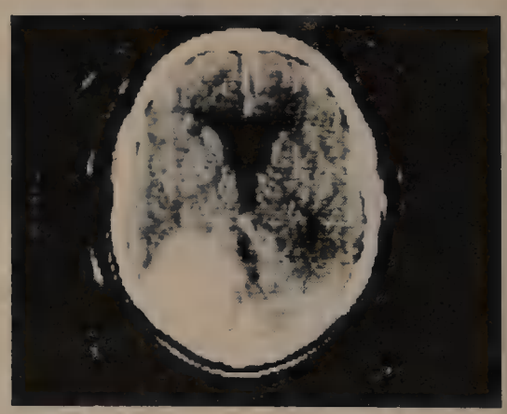
Echocardiograms from patients with aortic regurgitation.



Top curve represents survival in chronic hepatitis patients treated with steroids.



CT scan reveals large meningioma and mild hydrocephalus.



SCIENTIFIC AMERICAN MEDICINE
415 Madison Avenue, New York, New York 10017

Please enroll me as a subscriber to SCIENTIFIC AMERICAN *Medicine*. On receipt of this coupon you will send me the advanced two-volume text described in your announcement and update it regularly by sending me new monthly subsections. I understand that the annual subscription of $220 for SCIENTIFIC AMERICAN *Medicine* is tax deductible. If I am not entirely satisfied, I may cancel at any time during the first 60 days, returning all materials for a *complete refund*.

☐ Please enter my subscription for SCIENTIFIC AMERICAN *Medicine*
☐ I shall also enroll in the CME program
☐ I enclose a check made out to SCIENTIFIC AMERICAN *Medicine* for $220*
☐ Bill me ☐ VISA ☐ MasterCard

Expiration Date_____ Account Number_____

*Please add sales tax for California, Illinois, Massachusetts, Michigan, Ohio and New York

Name_____ MD Specialty_____

Address_____

City_____ State_____ Zip_____

Signature_____

Please allow six to eight weeks for delivery. All payments must be in U.S. dollars. CME is available outside the U.S. and its possessions for an additional $65. Surface routes used unless airmail delivery is prepaid. 7E

You may also order your trial copy by calling toll-free 1-800-345-8112 (in Pennsylvania, 1-800-662-2444).

*As an organization accredited for continuing medical education, the Stanford University School of Medicine designates this continuing medical education activity as meeting the criteria for 32 credit hours in Category 1 for Educational Materials for the Physician's Recognition Award of the American Medical Association, provided it has been completed according to instructions.

This program has been reviewed and is acceptable for 32 prescribed hours by the American Academy of Family Physicians.

This program has been approved by the American College of Emergency Physicians for 32 hours of ACEP Category 1 credit.

ing Russian cruise missiles the Russian negotiators insisted that any limits apply only to missiles with a range of more than 600 kilometers.) The issue took on added urgency with the discussion among the member states of the North Atlantic Treaty Organization (NATO) of the possible deployment of ground-launched versions of such systems in Europe, together with President Carter's decision to cancel the B-1 bomber and to emphasize the deployment of air-launched cruise missiles (ALCM's) on bombers based in the continental U.S.

In the case of long-range ALCM's the principal issue was not a ban but how the missiles were to be counted. Was a bomber carrying ALCM's to be counted as a single delivery vehicle no matter how many missiles it carried (as is the rule for a bomber carrying bombs or short-range cruise missiles) or was each ALCM to be counted as an individual delivery vehicle (as in the case of submarine-launched ballistic missiles)? The U.S.S.R. finally agreed to a complex compromise counting only the bombers, with the provisos that no existing type of bomber could be equipped with more than 20 cruise missiles and that the average number of cruise missiles on all bombers so equipped could not exceed 28. In addition if the number of bombers equipped with cruise missiles were to exceed 120, then the number of MIRVed missiles would have to be correspondingly reduced.

With regard to ground-launched and sea-launched cruise missiles the U.S. sought to have no limits applied, because other intermediate-range systems (such as the new Russian SS-20 missile) were not being limited in the agreement; the U.S.S.R. wanted a total ban. In the end testing and development were permitted, but deployment was suspended until the end of 1981 in order to allow further negotiations.

Eventually all these issues were resolved to the satisfaction of the two presidents and most of their advisers, but they continued to provide the basis for opposition to the treaty within the U.S. body politic, and as a result the ratification process was delayed. Ultimately external events—the matter of the "Russian Brigade" in Cuba, the hostage crisis in Iran and the Russian intervention in Afghanistan—overtook the process and brought the attempt to ratify the treaty to a full stop. Carter Administration officials involved in the effort (including myself) remain convinced that for all the troubles the treaty had on Capitol Hill, it would in the end have been ratified but for the Russian intervention in Afghanistan.

During the presidential campaign of 1980 Ronald Reagan and his supporters severely criticized SALT II, calling it "fatally flawed," sometimes on the basis of the issues I have cited. Immediately after the election President Reagan withdrew the SALT II Treaty from any further consideration by the Senate, but he declared the U.S. would abide by it as long as the U.S.S.R. did, and he initiated a new round of negotiations, called START (Strategic Arms Reduction Talks), as a means for generating a treaty more to his liking and that of his supporters. He also placed special emphasis on the need for major reductions in the number of deployed systems, in particular those that have very short flight times (both land-based and sea-based ballistic missiles) and are therefore felt to be particularly destabilizing.

Specifically, it has been reported that the opening U.S. negotiating position at START called for an overall reduction in the number of ballistic delivery systems by more than 50 percent (to 850 ballistic missiles bearing not more than 5,000 reentry vehicles altogether) and also a reduction of 50 percent in heavy ICBM's. In addition the U.S. position is said to be that consideration of limitations on bombers and cruise missiles must be deferred until a later time. It has also been reported that the original position of the U.S.S.R. called for a reduction in the number of delivery vehicles of all types (including bombers) to 1,800 for each side, a figure first proposed by President Carter in 1977 and forcefully rejected by the U.S.S.R. at that time. Recently President Reagan has said he would be more flexible on some of these issues, and he has specifically stated that the total number of ballistic missiles could be somewhere between the American and Russian opening figures.

From the point of view of the U.S.S.R. the main fault in the U.S. position is that it places sole emphasis on those deployments in which the Russians have an advantage and ignores those in which the U.S. has an advantage. From the American point of view the main fault in the Russian position is that it apparently continues to ignore the very important heavy-missile problem.

In addition to the issues I have been describing, which have for some years formed the basis for controversy both within the U.S. national-security community and between the U.S. and the U.S.S.R., two other issues have become increasingly important. One is the issue of mobile missiles, which has arisen mainly because of some of the proposed deployment modes for the MX, the projected successor to the Minuteman ICBM. The other is the matter of "reload capability." It is easy to count missile silos, but it is not easy to count the missiles themselves. Therefore if either side were to build and hold in reserve a number of missiles comparable to the number of silos, a new and serious source of potential instability would result. It remains to be seen whether the current situation will become deadlocked or whether there will be progress in overcoming these problems, but I for one am not optimistic.

There has been a great deal of public discussion in the U.S. of American concessions—real and imagined—made in the course of postwar arms-control negotiations with the U.S.S.R. Western folk history, however, contains virtually nothing about Russian concessions, even though the U.S.S.R. has made a number of important ones. One such concession has to do with the definition of the term "strategic delivery system." It has always been the U.S. position that a strategic delivery system is one that is deployed either in its home territory or at sea and that can reach the home territory of the other country from that deployment site. On the other hand, the Russian position has always been that a strategic delivery system is a system that is able to reach either home territory from its deployment site no matter where the site is.

This difference in viewpoint arises from the very different geopolitical situations of the two superpowers, and it shows up most sharply in the case of "forward-based systems," that is, those medium-range U.S. systems that are currently deployed in Europe and that can reach the U.S.S.R. only when so deployed. The U.S. has consistently refused to count such systems in the SALT totals, and the U.S.S.R. has consistently contended that they should be included. So far the Russians have accepted the American view in this matter. (Such systems are evidently now included in the separately organized talks in Geneva on intermediate-range weapons.)

A second Russian concession involves the British and French nuclear forces. The U.S.S.R. contends that the long-range delivery systems of these two NATO countries should be included in the SALT totals. The U.S. contends that SALT is strictly bilateral and so only the forces of the U.S. and the U.S.S.R. should be counted. In the SALT I agreement the Russians said they would accept the U.S. position provided that the British and French together did not deploy more than nine ballistic-missile submarines (more precisely, that the total number of submarines deployed by the U.S., Britain and France not exceed 50). The U.S. refused to acknowledge even this limitation, but the matter remains moot because that number has not been exceeded.

A third major concession by the U.S.S.R. involves the procedure for counting ALCM's on U.S. bombers. As long as the total number of these cruise



The 757 is the most fuel-efficient jetliner in the sky. In passenger comfort, it's superior to any other plane its size. Here's the perfect match of advanced technology and passenger conveniences. It's one way Boeing is helping to keep air fares one of the world's best travel values.

BOEING
Getting people together.

¡Speak Spanish like a diplomat!

What sort of people need to learn a foreign language as quickly and effectively as possible? *Foreign service personnel,* that's who. Members of America's diplomatic corps are assigned to U.S. embassies abroad, where they must be able to converse fluently in every situation.

Now you can learn to speak Spanish just as these diplomatic personnel do — with the Foreign Service Institute's Programmatic Spanish Course. You'll learn Latin American Spanish recorded by native speakers.

The U.S. Department of State has spent thousands of dollars developing this course. It's by far the most effective way to learn Spanish at your own convenience and at your own pace.

The course consists of a series of cassettes and accompanying textbook. Simply follow the spoken and written instructions, listening and repeating. By the end of the course, you'll be learning and speaking entirely in Spanish!

This course turns your cassette player into a "teaching machine." With its unique "programmatic" learning method, you set your own pace — testing yourself, correcting errors, reinforcing accurate responses.

The FSI's Programmatic Spanish Course comes in two volumes,each shipped in a handsome library binder. Order either, or save 10% by ordering both:

☐ **Volume I: Basic.** 12 cassettes (17 hr.), manual, and 464-p. text, $125.

☐ **Volume II: Intermediate.** 8 cassettes (12 hr.), manual, and 614-p. text, $110.

(Conn. and N.Y. residents add sales tax)

TO ORDER BY PHONE, PLEASE CALL TOLL-FREE NUMBER: **1-800-243-1234.**

To order by mail, clip this ad and send with your name and address, and a check or money order — or charge to your credit card (AmEx, VISA, MasterCard, Diners) by enclosing card number, expiration date, and your signature.

The Foreign Service Institute's Spanish course is unconditionally guaranteed. Try it for three weeks. If you're not convinced it's the fastest, easiest, most painless way to learn Spanish, return it and we'll refund every penny you paid. Order today!

112 courses in 35 other languages also available. Write us for free catalog. Our 12th year.



Audio-Forum
Suite 41M
On-the-Green,
Guilford, CT 06437
(203) 453-9794

AUDIO-FORUM®

Or visit our New York sales office: 145 E. 49th St., New York, N.Y. 10017 (212)753-1783

missiles remains less than 3,000 and certain counting rules are satisfied, only the number of bombers so equipped, but not the number of cruise missiles, need be counted and limited under the SALT ceilings. As in the case of the much more widely reported U.S. concessions, the U.S.S.R. has apparently made these concessions and others in a serious spirit of compromise in order to produce a mutually beneficial result.

Just a few weeks before the 1980 presidential elections the U.S. and the U.S.S.R. initiated negotiations designed to limit or eliminate intermediate-range missiles in Europe (technically known in the West as long-range theater nuclear forces). These negotiations are still in progress. Both general and particular factors contributed to the negotiations. The general factor is that there are a great many nuclear weapons (10,000 or more) deployed in or aimed at Western Europe, and yet very few of them are included in the START negotiations. The particular factor arises out of the Russian deployment of the SS-20, the first new weapon in its class in about 20 years. The deployment of the SS-20 began in 1977 and has been proceeding steadily ever since. By 1978 the situation had reached the point where many people in the Western countries became convinced that some kind of reply was necessary, at least for political reasons if not for strictly military ones.

At about this time Chancellor Schmidt of the Federal Republic of Germany made a speech in which he focused special attention on the situation; he asserted there was a need for some sort of highly visible land-based NATO system that would roughly counterbalance the SS-20. In December, 1978, in response to this growing concern, the NATO Council adopted what has become known as the "two-track" approach. One track called for the development and deployment of such a NATO system, and the other called for negotiations designed to make such a system unnecessary. One result of this dual approach was the decision by the U.S. and its NATO allies to deploy on land in Europe 572 intermediate-range missiles of two new types: the ground-launched cruise missile and the Pershing II, a modernized, longer-range version of the Pershing IA missile. As of this writing, neither has been deployed but the latter is expected to be later this year.

The other result of the two-track approach was the initiation of strictly bilateral talks between the two superpowers on this subject in Geneva in late October, 1980. The talks had hardly begun when the U.S. election results came in and both governments knew they would have to wait until the dust settled before anything could be accomplished.

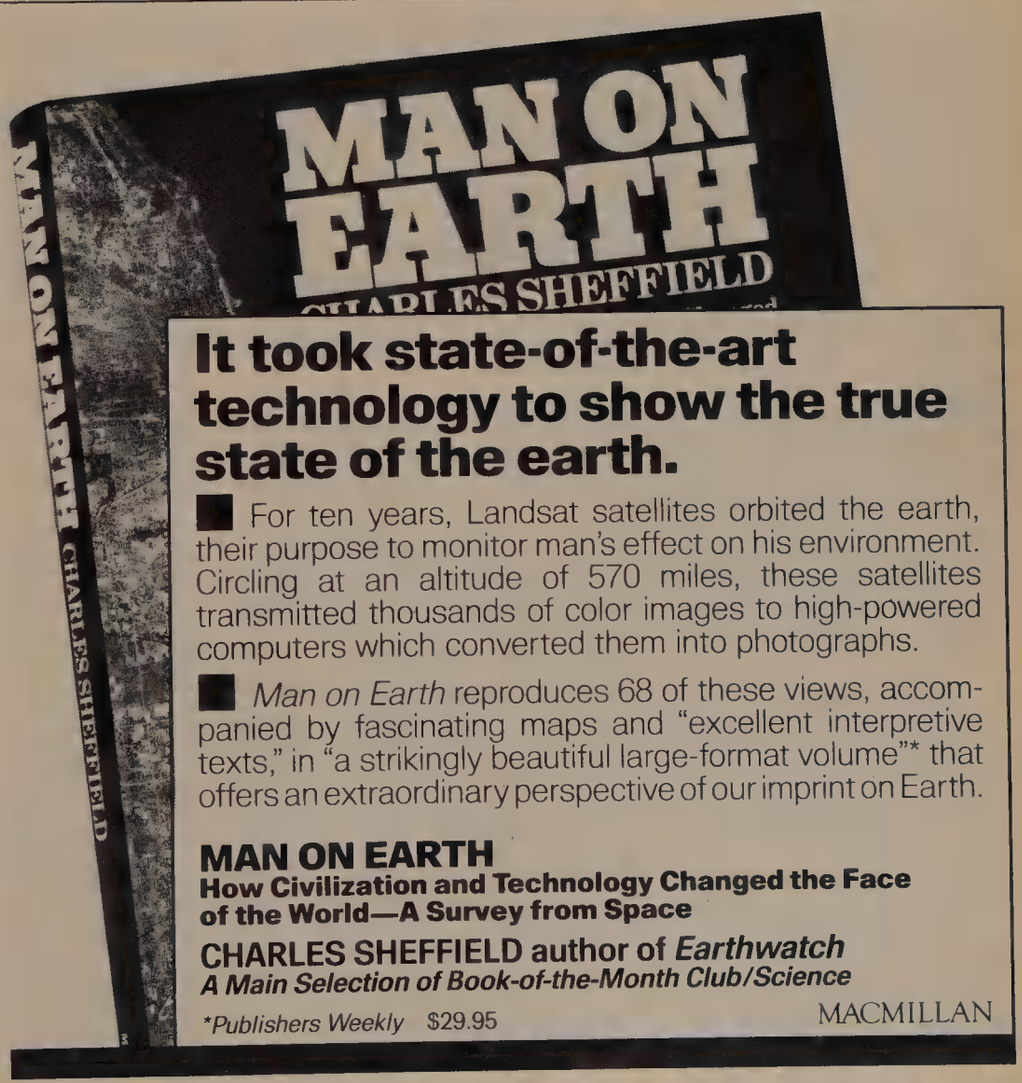
After some months of study the Reagan Administration resumed the negotiations and proposed what is known as the "zero-zero option." This proposal called for the elimination of all Russian medium-range missiles, including the SS-20's (or at least those west of the Urals), in return for a promise by the Western countries to forgo their plan to deploy a counterbalancing force. The response of the U.S.S.R., coming early in the new Andropov administration, proposed that the U.S. eliminate its plans to deploy any such missiles, and that the Russians reduce their SS-20's to a number (162) that would match the corresponding forces belonging to Britain and France.

Considered abstractly, the zero-zero option is clearly a desirable and reasonable proposal. It eliminates an entire class of weapons, one that may have filled a real gap in the 1950's but whose mission can be readily accomplished by other weapons each side now has in abundance. Considered concretely, however, the U.S. proposal calls for the U.S.S.R. to give up weapons it has already bought and paid for, whereas the proposal calls for the U.S. merely to forgo certain of its future plans. Therefore, desirable and reasonable as the zero-zero option may seem, in the current very bad political climate it will probably prove impossible to overcome the difficulty stemming from the asymmetry in the concrete situation. As of this writing each party has stuck with its original position, although President Reagan has indicated there might be some room for compromise on the American side.

From 1958 until about 1977 the principal method used by the U.S. to inhibit the development and deployment of devices suitable for conducting warfare in or from outer space was to practice moderation in its own programs and to urge the U.S.S.R. to do the same. Throughout that period the U.S. deployed space systems that performed important support functions (such as reconnaissance, surveillance and communications), and the military gradually developed an increasing reliance on such systems. Proposals for developing a general-purpose antisatellite (ASAT) system were frequently made in the U.S. by military and industrial organizations, but they were turned down by higher authorities. The basic argument was that the nation's space assets, as they are called, had become so valuable that a world in which neither superpower had antisatellite weapons was better than one in which both did, and that the U.S. therefore should not set a precedent by pushing ahead with the development of an ASAT system of its own. There was only one minor exception to this policy, and that was the development and deployment of an ASAT system of very



Passengers rate the Boeing 767 as a superb flying experience. It has two wide aisles and is the only wide-body with seven-abreast seating, so nearly everyone has either a window or aisle seat. Pilots rate the 21st Century flight control system as the most advanced in the world. BOEING
Getting people together.





World travelers choose the 737 because it has one of the best on-time records. The 737-300 is the latest addition to the family. This new jetliner, with added passenger amenities, advanced flight controls and quieter engines, will make flying an even better travel experience.

BOEING
Getting people together.

Carnegie-Mellon University

Program for Technical Managers

March 4-23, 1984

The greater your management responsibilities, the less your contact with actual technical work. Yet building an effective staff of younger professionals and setting effective technical strategies demands a familiarity with the latest developments in modern science and engineering and their potential applications.

In response, Carnegie-Mellon University has developed the *Program for Technical Managers*. Designed for experienced technical managers, the Program is an intensive technology update coupled with realistic insights of future trends and applications. In addition, auxiliary topics such as law, sociological implications, economic considerations, entrepreneurial techniques and product liability concerns are presented.

Expand your capacity to help solve your organization's technical problems while increasing productivity. Make time for the *Program for Technical Managers*. For your organization. For the next 20 years of your career.

Call Frank E. Nowak (412/578-2207) for information. Or write Post College Professional Education, Carnegie-Mellon University, 405 MMCH, Pittsburgh, PA 15213.

limited capability on Johnston Island in the early 1960's because the U.S. suspected the U.S.S.R. was developing a system that involved stationing extremely powerful nuclear weapons in orbit around the earth. The Russian ASAT development never did come, however, and after only a few years the U.S. decommissioned its specialized ASAT system.

In spite of U.S. restraint the U.S.S.R. started in 1967 to test a general-purpose ASAT system of limited capability. Evidently the Russians gave this development a lower priority than most parts of their space program, and it proceeded rather slowly. On occasion the U.S. pointed out that it was deliberately practicing moderation in this area and urged the U.S.S.R. to do the same. That may have had some effect, but the Russian program continued nonetheless.

Finally in 1977 the Carter Administration decided that something more specific had to be done by way of response to the Russian ASAT development, and a three-pronged program was initiated. One element of the program was the decision to begin a full-scale effort to develop an American ASAT system. (A contingent development had already been started during the Ford Administration.) A second was a plan to explore and develop means for defending American satellites against a possible attack by Russian ASAT weapons. The third was the initiation of negotiations designed to forestall further development and deployment of ASAT weapons by either side, if possible.

The first formal negotiations on ASAT systems took place in Helsinki in 1978. Two further negotiating sessions were held in 1979, after which the negotiations were adjourned indefinitely; to this date they have not been resumed. There were both internal and external reasons for the failure to accomplish anything. Internally the problem was twofold. First, the U.S. Government was unable to reach a unified position with respect to what ought to be done. On the one hand, top officials of the Air Force and the Department of Defense placed such a high value on the nation's existing space assets that they were willing to forgo development of an ASAT system if this would ensure that the Russians also did so. The position of the Navy, on the other hand, was that Russian ocean-surveillance satellites presented such a serious threat to the U.S. fleet that the nation needed an ASAT system of its own no matter what the Russians did. Furthermore, officers in the middle levels of the Air Force did not like having any artificial "political" barriers placed in the path of space developments of any kind. On top of these differences the fact that the U.S.S.R. had

already conducted more than a dozen ASAT tests in the preceding decade raised strong pressures for at least an equivalent number of similar tests by the U.S.

The second problem internal to the negotiations was that the Russians also seemed to be ambivalent about the issue. In the negotiations they indicated that in their view there could be serious and valid reasons for possessing an ASAT capability, and they also raised a series of questions about the U.S. space shuttle and what role it might have in this connection. In addition to all these problems relating directly to the substance of the ASAT negotiations, the same general deterioration of the international climate that interfered with SALT II and the comprehensive-test-ban negotiations also operated as a factor strongly inhibiting further progress in this area.

In addition to these so far futile attempts to forestall the development of antisatellite systems, the U.S. and the U.S.S.R. have successfully negotiated more limited agreements in the area. The most important are the clauses in the Limited Test Ban Treaty prohibiting nuclear explosions in outer space, the Outer Space Treaty of 1967 and the clauses of the SALT I Treaty that legitimize and protect reconnaissance satellites used for monitoring arms-control agreements and that prohibit the development and deployment of space-based ABM systems. The Outer Space Treaty prohibits stationing "weapons of mass destruction" in outer space or on celestial bodies. The phrase in quotation marks is always interpreted as including nuclear weapons and excluding non-nuclear ones.

More recently, in 1981, a proposal for an agreement designed to forestall preparations for warfare in space was placed before the United Nations by President Brezhnev. In brief, the proposal called for a prohibition on the deployment of weapons of any kind—not just weapons of mass destruction—in orbit around the earth, on celestial bodies "or in outer space in any other way." The proposal seems carefully to avoid mentioning ground-based ASAT systems of the type both nations are developing. Except for the Brezhnev proposal in 1981 nothing further of a formal nature has happened in this area since the adjournment of the ASAT negotiations in 1979.

There has been much talk lately about the possibility of orbiting "directed-energy weapons" in space. This talk has usually involved three classes of weapons: conventional optical lasers, particle-beam devices and a new category of systems, cryptically referred to by their advocates as "third-generation nuclear weapons," that is known to include X-ray lasers powered by nuclear explosions. These three very different classes of weapons are in very different stages of research and development. The first class, optical lasers, is by far the most advanced. No such system has yet been designed in detail, but presumably one would consist of an assemblage of subsystems some of which would be similar to, although bigger or more capable than, devices that have already been operated in laboratories on earth.

A laser-weapon system that could destroy "cooperative" targets could probably be deployed in space before the end of the century, although at great cost. (I am referring here to high-energy systems that physically destroy their target, not to low-power systems that blind optical sensors or otherwise cause limited special damage.) The two other classes of directed-energy weapons are still only in the research stage. No systems development for either one has yet been carried out or is currently planned, and in my judgment there is no chance whatever that a practical operational device of even limited capability will be deployed in space during this century.

Although the development of such systems is not now prohibited by any international agreement, their development is strongly inhibited by an even more powerful factor: their extremely high cost. The best estimates available indicate that the development cost alone of a "laser battle station" would be about 10 times as much as that of other roughly similar modern high-technology systems (that is, tens of billions of dollars as opposed to billions of dollars), and that the deployment of a system capable of intercepting more than a small fraction of the Russian strategic force would cost a large fraction of a trillion dollars. These huge costs have, so far at least, boggled the minds of responsible officials, and in spite of President Reagan's "Star Wars" initiative there are still no serious plans for the full-scale development of any such battle stations based on any of the various hypothetical types of "killer beams."

The U.S. and the U.S.S.R. each have unique characteristics that make them particularly difficult as negotiating partners when they are dealing either with each other or with third parties. In the case of the U.S. the most serious of these special difficulties arise from certain fundamental mechanisms of the American system of government. These are the requirement for a two-thirds vote in the U.S. Senate for ratification of a treaty and the exceptionally long presidential campaign that the nation goes through every four years.

It was the requirement for a two-thirds vote in the Senate that kept the U.S. out of the League of Nations after World War I, prevented the U.S. from acceding to the Geneva Protocol on chemical warfare of 1925 until fully 50 years had passed and more recently has prevented the U.S. from ratifying the Threshhold Test Ban Treaty, the Peaceful Nuclear Explosions Treaty and the SALT II Treaty. It seems likely that each of these treaties would have been ratified in a timely fashion if only a simple majority were required. Indeed, this particularly difficult and awkward situation is the purpose of the rule. The idea of the country's founding fathers was that the U.S. should avoid "foreign entanglements." When the country was actually isolated by two great oceans, that probably made good sense. In a world that is tightly integrated by high technology and in which Russian missiles in Siberia are less than a half hour away from their targets in the U.S., however, it is a drastically different matter.

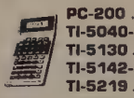
The problem is not just one of failing to ratify certain treaties. More important, I think, is the effect this factor has on the negotiating process. Every president early in his term discovers that arms control is even more controversial than he thought it was, and that the ultimate measure of the acceptability of what he negotiates is getting the two-thirds vote in the Senate. Accordingly arms-control policy is developed not with an eye to a national consensus or even to a majority of the Congress, but rather with the objective of somehow capturing the support of those five or six senators who are two-thirds of the way over toward the "no" end of the political spectrum. For good or ill this factor has greatly colored the policies of every president from Kennedy to Carter, and the negotiating instructions to our delegations overseas have always been characterized by extreme caution and conservatism. Contrary to the common belief, this often means that as time goes on in a negotiation the U.S. position undergoes sudden changes, not in response to what the other side may have proposed but in a difficult and sometimes futile process of presidential maneuvering to please a few key senators and those members of the executive branch, particularly in the uniformed military, who have special influence with those senators.

The problem presented by the quadrennial presidential campaign is also serious. Because arms-control policy is intrinsically controversial, all new presidents soon discover that it is absolutely essential to consult all interested elements of the executive branch before they can finally establish arms-control policy and work out detailed negotiating instructions. This process typically takes up most or all of the first year of a new president's term. Later, in the fourth

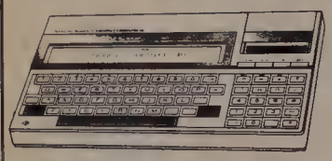
SAVE AT ELEK-TEK

TEXAS INSTRUMENTS

BA-2	33	PC-200	52
BA-35	22	TI-5040-2	49
TI-35-SP	17	TI-5130	60
BA-55	43	TI-5142-3	75
TI-55-2	33	TI-5219	130
TI-57	29	TI-5310	97
TI-66	52	TI-LCD-Prog	50



TEXAS INSTRUMENTS



$189

COMPACT COMPUTER 40

34K bytes of ROM; 6K bytes of RAM; Constant Memory Feature; Convenient plug-in applications software; Enhanced BASIC programming language; Easy-to-use typewriter-style (QWERTY) keyboard; Battery-powered; 31 character LCD display; Peripheral interface connection.

HP-75C 16K PORTABLE COMPUTER

$719 Mfr. Sugg. 995



The battery powered portable computer which includes a built-in HP-IL interface, card reader, and time and appointment modes. Typewriter keyboard.

82164A HPIL/RS-232C Interface 235
82168A HPIL Acoustic Modem 395
82169A HPIL/HP1B Interface 315
82700A 8K Mem. Module 150

NEW!

HEWLETT-PACKARD SLIMLINE Shirtpocket Styled Power Packed Programmable LCD PROBLEM SOLVERS

HP-10C Scientific *(New)* $54
HP-11C Scientific 70
HP-15C Scientific *(New)* 90
HP-12C Financial 90
HP-16C Programmer *(New)* ...90
HP-97 Desktop Scientific ..560

HP-41C $145
HP-41CV 200



Optical Wand.....95
Card Reader.....145
Printer 283
Quad R.A.M.(for HP41C)..**60**
Ext. Memory Module....60
Ext. Function Module....60
Time Module.....60

	Mfr Sugg	Elek-Tek
HP-IL Loop Module	125	95
Digital Cassette Drive	450	335
Printer/Plotter(HP-IL)	450	335

50%-60% off orig. mfr. suggested prices

HP-32E Scientific 32.50
HP-34C Scientific 50.00
HP-37E Financial 32.50

HEWLETT PACKARD | PERSONAL COMPUTATION DEALER

CALL TOLL FREE 800-621-1269
EXCEPT Illinois, Alaska, Hawaii

Corp. Accts. Invited. Mastercard or Visa by mail or phone. Mail Cashier's Check, Mon. Ord. Pers. Check (2wks to clr.) Add $4.00 1st item. (**AK, HI, P.R., Canada add $10.00 first item**) $1.00 ea. add'l shpg. & handl. Shipments to IL address add 6% tax. Prices subj. to change. **WRITE for free catalog. 30 day return policy for defective replacements only. ALL ELEK-TEK MERCHANDISE IS BRAND NEW, FIRST QUALITY AND COMPLETE.**

ELEK-TEK, inc.
6557 N. Lincoln Ave., Chicago, IL 60645
(312) 631-7800 (312) 677-7660

year of his term, the president finds himself faced with a primary campaign within his own party, followed by the general campaign in which he must face the candidate of the opposing party.

The problems created by the exceptionally long American presidential campaign become evident when one considers the problems that first President Johnson and then Vice-President Humphrey had with this kind of issue in 1968, the problems that President Ford had over the SALT II negotiations and the Panama Canal negotiations during his primary campaign against Ronald Reagan in 1976 and the problems that President Carter had first with Senator Kennedy and then with Ronald Reagan in 1980 (these last being of course greatly exacerbated by the events in Iran and Afghanistan). In sum the U.S., at least in this era of four-year presidencies, is only in a position to negotiate seriously for two years of every four and even then in a way that forces the president to focus special attention on those few people who are two-thirds of the way over toward a "no" vote.

In addition to these peculiarly American problems the bureaucratic difficulties that plague all complex societies also have a strong negative influence on the ability of the U.S. to negotiate. This is primarily because the "nay-sayers" have only to stop or slow the process in any way at all in order to achieve their goals, whereas those who want to help the president achieve his arms-control goals must find a genuinely suitable pathway through what is often uncharted territory.

Perhaps the most serious problems with the Russians as negotiating partners are their penchant for secrecy and their tightly controlled political system. These two problems together are at the root of the difficulties Americans always have in dealing with the verification problem. The U.S. has extraordinarily powerful and sophisticated "national technical means" of verification, including observation satellites, seismic stations all over the world and other ways of eavesdropping. With these the U.S. can and does learn a great deal about what is going on in the U.S.S.R., and most people who are closely acquainted with these technical means and their output hold them to be adequate for verifying all the treaties that have been signed so far. Nevertheless, it is necessary and prudent for the U.S. Government, meaning particularly the intelligence community, to classify the details of these systems and to keep secret the evaluations of their performance.

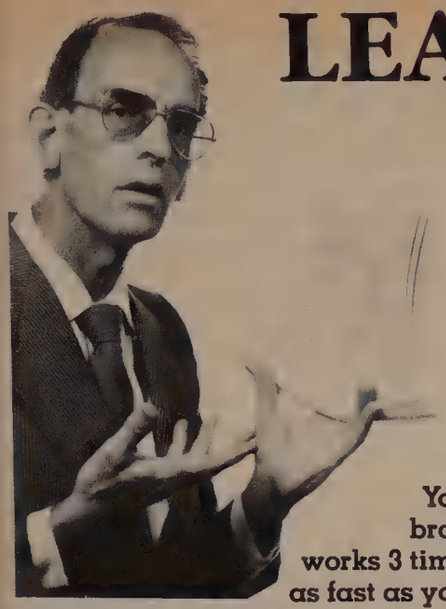
Given this genuine need for secrecy on the part of the U.S. and given the fact that there are always some people inside the system who are not fully satisfied with the performance of the national technical means of verification, it should come as no surprise that those officials and other members of the American body politic who hold extremely negative views of the U.S.S.R. in general and take a cautious approach to arms control in particular are doubtful about the ability of the technical means to uncover every deception that might be invented. The U.S.S.R. is, after all, a huge country with large areas entirely off-limits to foreigners, and it is not just paranoia to suppose that it might be doing something in such places that is significant but not known to the U.S. intelligence community.

The political control that extends throughout all levels of life in the U.S.S.R. makes the problem much worse. It is trite but true to say there are no organized opposition parties, no public-interest groups, no lobbies and no "whistle blowers" of the kind that in most other industrial societies are capable of uncovering and publicizing government misdeeds, including treaty violations. In my discussion of the test ban I mentioned that one of the general problems that has persisted throughout the entire postwar period of bilateral negotiations is that the Americans always ask for more intrusive means of verification than the Russians are willing even to discuss, much less accept, and the Russians are always complaining that the Americans persist in trying to spy on them, to undermine their sovereignty and to otherwise interfere in their internal affairs.

This problem will, I believe, continue into the indefinite future, and important elements of the American body politic will continue to harbor grave doubts about the adequacy of any conceivable verification system. I see no easy way of dealing with this difficulty until and unless there are substantial changes in the political behavior of the U.S.S.R., and I see no reason to expect that in the foreseeable future.

The situation is not totally bleak. The views of the Russian leaders on matters of secrecy and sovereignty have been evolving, and they have relaxed some of their earlier positions on related questions. For example, their acceptance of reconnaissance by satellite, their agreement in principle to exchange geophysical data in connection with the Threshold Test Ban Treaty, their acceptance of certain special on-site activities by foreign observers in the Peaceful Nuclear Explosions Treaty and their acceptance of special seismic stations and a carefully hedged system of voluntary on-site inspections in the comprehensive-test-ban negotiations all attest to this evolution. Whether it is proceeding at a pace that can keep up with American suspicions remains to be seen.

LEARN MATH — *EASY!*



Your brain works 3 times as fast as your fingers can write! That's where the Calculator comes in — you can go three times as fast with your machine doing the dirty work.

It's easier and more interesting. And you'll understand it better when it's fun. (If you haven't yet tried out a modern scientific-statistical calculator, you're in for a pleasant surprise — see below!)

WHY MATH? To get a promotion or a new job. Just to hang on to that job you've got. To quicken your mind. To understand this world you live in. To impress people. Just for some fun you'll feel good about. (Read those letters below to see how it has worked for others.)

Do it at home... You can start right away and go as fast as you like! Put it aside while you're busy with something else — in short, *set your own pace!*

Lots of us like to avoid being compared with the other members of a class — we like to set our own standards. Too, we like to save the time and expense of going to a school or seminar. (Best of all, it's a lot cheaper to study at home where the rent is already paid — *read on!*)

Let's be honest. You wouldn't expect to learn to play baseball or chess by watching an expert do it. Math is the same — you'll need to do some work. BUT — with my calculator method, you'll see it's a whole new ball game!

THIS METHOD IS FRIENDLY! You can use it on your own. You follow interesting examples step-by-step with calculator keystrokes — instead of algebra.

Let me introduce myself — I'm Dr. George McCarty and I teach at the University of California. Tens of thousands of people like yourself have already used my guidebook to teach themselves calculus. Recently I have teamed up with another professor. We put those same tested and proven techniques to work for you in a new *Statistics Guidebook.*

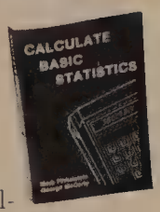
Even if you've had trouble with math, this method will work for you. Choose the one you want to learn, calculus or statistics (or both) and start *today!*

As pennywise Ben Franklin said, "An investment in knowledge pays the best dividends."

CHOOSE STATISTICS

It's used everywhere — to set insurance rates and to assess a drug's effectiveness, to detect a crooked roulette wheel and to measure lightbulb lifetimes, to audit the books and to find your position at sea.

People have always known that their experience was not a perfect guide to the future. Statistics shows us that uncertainty is not only inevitable but it itself is measurable.



You'll find out about this in my guidebook. You'll read informal explanations with detailed calculations for topics like "confidence intervals" and "contingency tables."

YOU'LL BE DOING Inference and Hypothesis Testing, the Chi-Square Distribution, ANOVA, Regression and Correlation, Non-Parametric Methods — they're all here.

SUPPOSE YOU HAVE ALREADY STUDIED STATISTICS. You'll use Quick-Reference Guides to go straight to a topic you want to refresh yourself on — then get *right-now* answers in your own field, business, biology, science, psychology, education.

The Guidebook is guaranteed to work with any calculator or computer. If you would like to start out with a fancy new machine, order my Statistics Kit which includes the guidebook and a TI-35 calculator which finds means and standard deviations at the touch of a button. It has an LCD display, 1000 hour batteries — the works.



CHOOSE CALCULUS

The math of science and engineering — satellite orbits and compound interest, bacterial growth and radioactive decay, marginal profits and reaction rates.



You'll find out about these concepts when you use the Calculus Kit. This Kit contains my guidebook and a TI-35 calculator which finds log and trig values at the touch of a button. (It's modern, too, with LCD display and 1000 hour batteries.)

In the guidebook you'll learn both differential calculus and integral calculus — *even differential equations!*

NEED TO EVALUATE functions, areas, volumes — solve equations — use curves, trig, polar coordinates — find limits for sequences and series? It's all here!

If you're in the biological, social or physical sciences, you'll be doing Bessel functions, carbon dating, Gompertz' growth curves, half-life, future value, marginal costs, motion, cooling, probability, pressure — and plenty more.

Important numerical algorithms are here, too: rational and Padé approximation, bracketing, continued fractions, Euler's method, Heun's method, iteration functions, Newton's method, predictor-corrector, successive substitutions, Simpson's method and synthetic division.

If you already have a scientific calculator, you can get just the *Calculus Guidebook.*

George McCarty

P.S. — YOU *CAN* DO IT! I'm certain I can show you this easy way to success with math, so *I guarantee it to you unconditionally* — if you are not satisfied just send it back for a prompt refund!

PHONE MC or VISA ORDERS TO
(714) 831-2637
Ask for our Catalog of books and calculators.

LOOK AT WHAT THEY SAY

- "...*excellent diagrams and clear explanations.*" —POPULAR SCIENCE
- Professor John Ball of Harvard College writes: *"I wish I had had as good a calculus course."*
- S.C. McCluney, Jr. of Philadelphia writes: *"With your book and a calculator the whole idea becomes clear in a moment, and is a MOST REFRESHING EXPERIENCE."*
- Professor H.I. Freedman of the U. of Alberta, states, *"There can be no question as to the usefulness of this book...lots of exercises...very clearly written...easy reading."*
- "...*straight-to-the-point explanations...*" —HP Key Notes
- Tektronix Engineer Bill Templeton says, *"The best, most clearly written book I have seen for improving your math skills."*

TAX DEDUCTIBLE FOR PROFESSIONALS

YES, I want to learn math **fast** and **easy** with your *new method.* If I am not more than satisfied I can return it for a quick refund. RUSH me the guidebook or complete kit I select below (kit contains the guidebook and a modern TI-35 calculator).

☐ Send **FREE** Catalog of books & calculators

STATISTICS GUIDEBOOK		KIT
How many?	☐ $14.95	☐ $39.95
Handling/Shipping	1.00	2.00
Tax (within Calif.)	.90	2.40
TOTAL		

CALCULUS GUIDEBOOK		KIT
How many?	☐ $19.95	☐ $44.95
Handling/Shipping	1.00	2.00
Tax (within Calif.)	1.20	2.70
TOTAL		

For AIR SHIPMENT add $3.00 per item

Phone (714)831-2637 (MC or VISA only)

Enclosed is ☐ Check, ☐ M.O.; or charge my MC/VISA... exp._____

Card #_____

Name _____

Address _____

City _____

State, Zip _____

OUTSIDE USA, **surface shipping is $2.00 per book or $5.00 per kit. Air shipment is $5.00 per book or $10.00 per kit. U.S. FUNDS PLEASE.**

EduCALC Publications, Dept. AO
27953 Cabot Road
South Laguna, CA 92677

©1982

THE AMATEUR SCIENTIST

Thinking about physics while scared to death (on a falling roller coaster)

by Jearl Walker

The rides in an amusement park not only are fun but also demonstrate principles of physics. Among them are rotational dynamics and energy conversion. I have been exploring the rides at Geauga Lake Amusement Park near Cleveland and have found that nearly every ride offers a memorable lesson.

To me the scariest rides at the park are the roller coasters. The Big Dipper is similar to many of the roller coasters that have thrilled passengers for most of this century. The cars are pulled by chain to the top of the highest hill along the track. Released from the chain as the front car begins its descent, the unpowered cars have almost no speed and only a small acceleration. As more cars get onto the downward slope the acceleration increases. It peaks when all the cars are headed downward. The peak value is the product of the acceleration generated by gravity and the sine of the slope of the track. A steeper descent generates a greater acceleration, but packing the coaster with heavier passengers does not.

When the coaster reaches the bottom of the valley and starts up the next hill, there is an instant when the cars are symmetrically distributed in the valley. The acceleration is zero. As more cars ascend, the coaster begins to slow, reaching its lowest speed just as it is symmetrically positioned at the top of the hill.

A roller coaster functions by means of transfers of energy. When the chain hauls the cars to the top of the first hill, it does work on the cars, endowing them with gravitational potential energy, the energy of a body in a gravitational field with respect to the distance of the body from some reference level such as the ground. As the cars descend into the first valley much of the stored energy is transferred into kinetic energy, the energy of motion.

If the loss of energy to friction and air drag is small, the total of the potential and kinetic energies must remain constant throughout the descent and even throughout the rest of the ride. The coaster gains kinetic energy and speed at the expense of potential energy. If the first valley is at ground level, the transfer is complete, and for a moment the coaster has all its energy in the form of kinetic energy.

Without energy losses the coaster could climb any number of hills as high as the one from which it is released (but no higher). To be sure, friction and air drag do remove energy from the coaster, and its total energy content dwindles. It can no longer climb high hills, which is why the last stages of the track consist only of low hills.

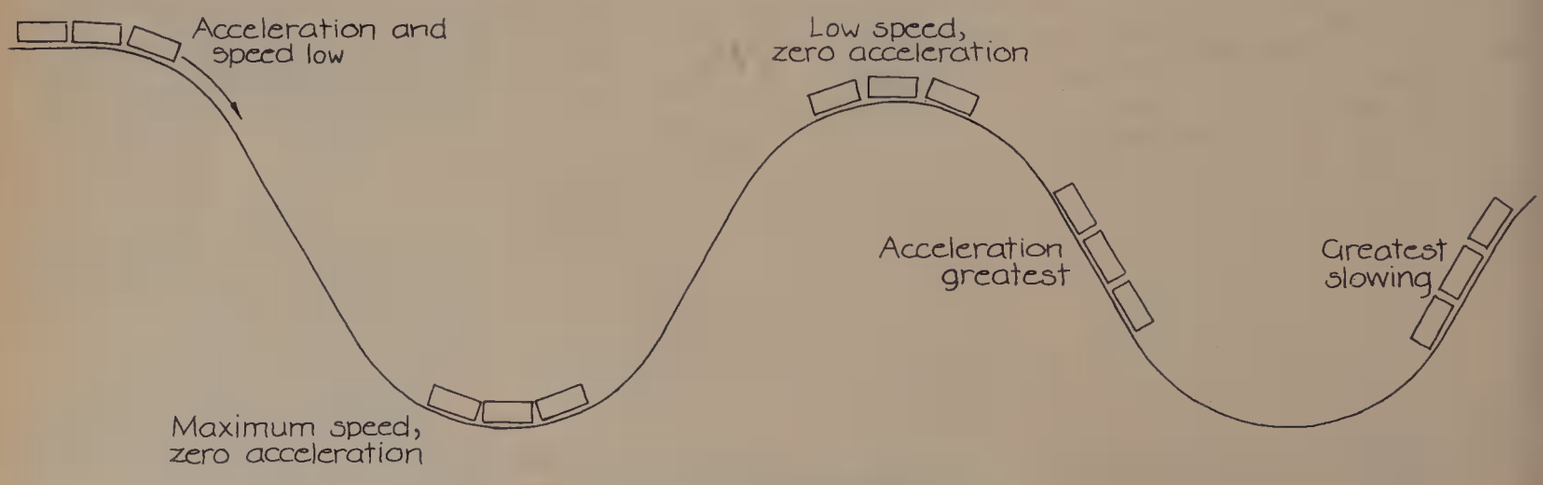
The length of a ride on a roller coaster depends on the speed. If the ride is to be fast, the launching hill should be high so that the total energy is large. The rest of the track should be low so that most of the energy remains kinetic.

The choice of a seat on a roller coaster makes a difference in the ride. Some people prefer the front seat because the descent from the launching site presents the pleasingly frightening illusion of falling over the edge of a cliff. Other people prefer the psychological security of the rear seat.

The choice of a seat also determines the forces felt by the passenger. Consider the first descent. The front car starts down slowly because little of the coaster's energy is then kinetic. The speed of the cars increases as an exponential function of time, so that the rear car starts down at a much higher speed than the front car did. Although the passengers in the front car get an unobstructed view of the descent, the passengers in the rear car have a stronger sense of being hurled over the edge.

At the edge one force on the passenger is from the change in the direction of his momentum vector. Initially the vector is horizontal, but soon it points toward the valley. The force necessary to effect this change in direction is delivered by the safety bar or seat belt that keeps the passenger in the car. That force, which points downward and back toward the hill, is part of the thrill of the ride. A passenger in the rear feels the force more than a passenger in the front because the size of the force is proportional to the momentum, which is greater for the passenger in the rear.

The story is different in the valley. Again a force from the coaster is necessary to redirect the passenger's momentum. This time the momentum is initially downward toward the bottom of the valley and then is redirected toward the top of the next hill. The front passenger has a large momentum and is subjected



The energetics of a roller coaster

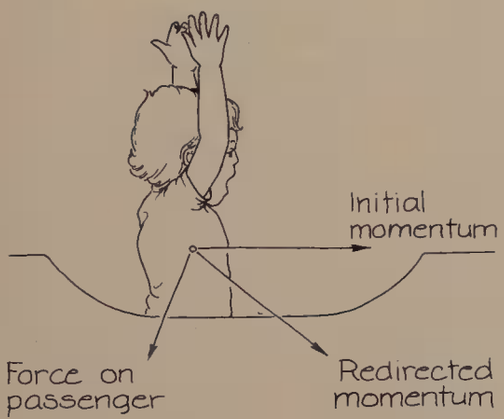
to a large force. By the time the rear car reaches the bottom of the valley the movement of the front cars up the next hill has slowed the coaster. A passenger at the rear has less momentum and is subjected to a smaller force.

At the crest of the hill the passenger gets a force leveling his momentum vector. At the rear of the coaster the force can be considerable if the front is already well down the next slope. To a passenger at the rear who is loosely held in place by a safety bar a fast trip over a hill provides a brief sensation of being lifted from the seat. He arrives at the crest with a large momentum. Until he encounters the safety bar and is redirected he continues to travel upward even though the coaster has leveled out below him. The faster the coaster goes over a hill, the greater the sensation of being thrown free.

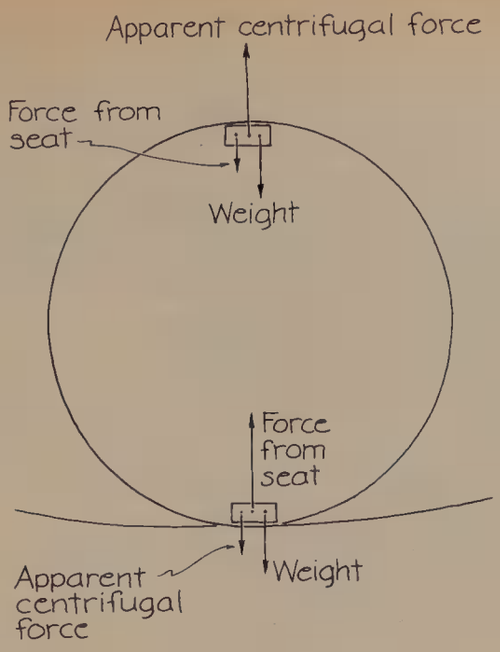
The brave passenger is one who rides the roller coaster without holding on. I tried this once while arriving at the crest of a hill at high speed. I avoided being thrown free of the coaster by catching my thighs on the safety bar at the last instant. Thereafter I kept a tight grip on the safety bar.

Roller coasters such as the Big Dipper have been around for more than 50 years. Recently a new type of coaster has appeared. The principles are seen in the Double Loop and the Corkscrew. The Double Loop at Geauga Lake begins like the Big Dipper in that a chain pulls the cars to the top of the first and highest hill. After the coaster travels over a few smaller hills (and before it loses too much of its energy to friction and air drag) it runs through two vertical loops. The ride is splendidly unnerving. In the times I managed to open my eyes while traveling through the loops I saw the world turn upside down, the ground race up toward me and the world turn upside down again.

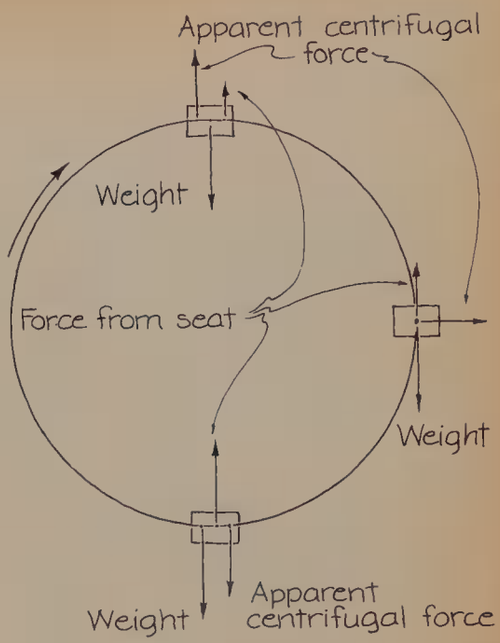
The coaster on the Double Loop is held onto the track by a double set of wheels, one set on the top of the rails and the other set on the bottom. When the coaster is on the normal section of the track, its weight rests on the top set of wheels. When it is in the loop, the other set of wheels can come into play. They keep the cars from flying off the track.



Forces that redirect momentum



Forces in a coaster that loops



What you feel in a Ferris wheel

When the coaster enters a loop, I sense three forces. One is my weight, which of course is directed downward. Another is the force from the seat. The third is the apparent centrifugal force downward, which seems to add to my weight; it makes me feel as though I am being pushed into the seat. At the top of a loop the apparent centrifugal force is upward, and I feel light.

The centrifugal force is a fiction. No outwardly directed force is at work. The notion of a centrifugal force is useful, however, since it easily explains what a passenger feels. The perspective of someone on the ground is more to the point: a combination of real forces causes the rider to move in a circle instead of a straight line.

If the circular motion is to be maintained, the net force must be toward the center of the circle. At the bottom of a loop the passenger's weight vector is downward (and therefore away from the center of the loop). An upward force acts on the passenger from the seat. Since the push from the seat is greater than his weight, the net force points toward the center of the loop and he begins the circular motion. From the passenger's perspective the large push from the seat is sensed as a centrifugal force pressing him into the seat.

At the top of the loop the forces have changed. The passenger's weight is still the same and is still pointed downward, that is, toward the center of the loop. The push from the seat is also downward. The two forces combine in the net force that makes the rider continue in the circle.

This time the force from the seat is smaller. One reason is that at the top of the loop the coaster has less kinetic energy and so is traveling slower. Moreover, the force from the seat is now augmented by the rider's weight vector instead of having to oppose it. The rider senses the force from the seat as a small centrifugal force.

How high must the coaster be at the start of its journey (with essentially no initial speed) if it is to have at the top of the loop the speed that will hold it firmly on the track? To answer the question I made two assumptions. The first was that the coaster has only one car. The second was that the energy losses from friction and air drag are negligible. With these assumptions I found that the first hill must be higher than the top of the loop by at least half the radius of the loop.

The first assumption is a convenient simplification. If the coaster is long, one must consider the rise and fall of its center of mass rather than considering only one car. Since only part of the coaster is at the top of the loop at any given instant, the center of mass never reaches that height, and so less energy is actually required to keep the coaster on the track than would be needed if there were only one car.

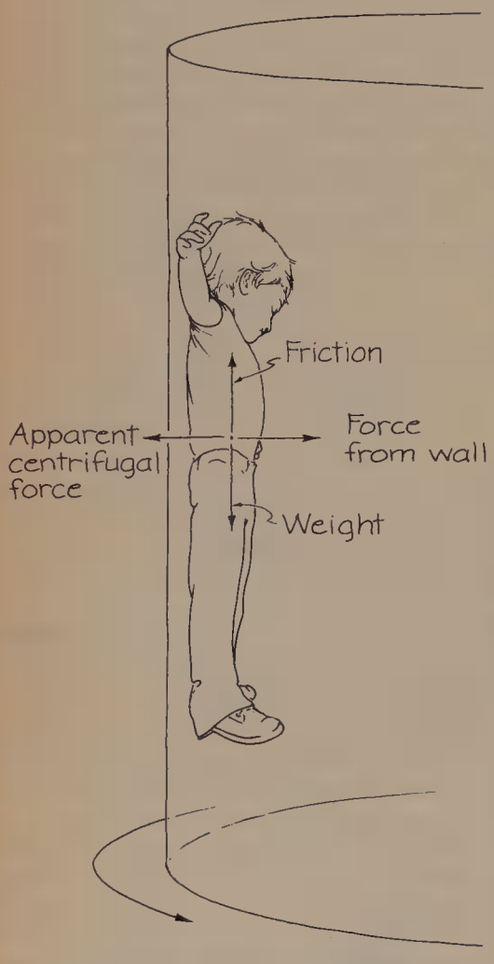
As for the second assumption, if the losses of energy were entirely negligible, the unpowered coaster would arrive at the loop with all the energy it got at the launch. The intervening hills and valleys would not matter. They do matter, of course, because they provide more opportunity for energy losses. Therefore the initial hill must be higher than the theory would indicate. On the Double Loop at Geauga Lake the initial hill is considerably higher than the theoretical value, so that the coaster is still traveling

at a good clip when it reaches the top of the loop.

The Corkscrew is a similar roller coaster except that the loops are helical. Once the coaster enters the loops it moves in a corkscrew fashion until it emerges again. At two points the passengers are fully upside down.

The physics of this ride is similar to that for the Double Loop. The major difference lies in the direction of the apparent centrifugal force. With the Double Loop the center of the motion in a loop is at a single point. The centrifugal force appears to be directed radially outward from that point. As the coaster travels around the loop, this force rotates in a vertical plane. With the Corkscrew the center of motion continuously moves vertically and horizontally as the coaster travels through the loops. Hence the direction of the apparent centrifugal force is not confined to a vertical plane. This added feature is one reason the Corkscrew has become so popular with coaster addicts.

Geauga Lake has two other rides that are similar to the standard roller coaster. The water slide starts high above the ground. Water pours down the interior of the slide to provide lubrication and even a small amount of propulsion. The principle is simple: the initial gravitational potential energy is steadily converted into kinetic energy, so that the slider's speed increases during the descent. The lubrication provided by the water diminishes the loss of energy to friction.



The role of friction in the Rotor

The other ride is the Gold Rush Log Flumes. Passengers board a small boat shaped like a hollow log. It is really a car like the ones on the Big Dipper. Water flowing through the flume pushes the boat along until it is engaged by a chain system that drags it up a tall hill. From the crest the boat descends rapidly down a slope of about 45 degrees. At the foot of the hill it speeds into a trough of water, which quickly slows the motion and satisfyingly drenches the passengers. They also seem to be thrown forward, but the experience is illusory; what happens is that they continue to move forward briefly at the former speed.

Most of the other rides at an amusement park are based on rotational mechanics. The mildest of them is the merry-go-round. Here the rate of rotation is just enough to give the passenger a moderate sensation of centrifugal force. He seems to be thrust outward. Actually his body leans outward because the horse moves away from him as it travels in a circle and ends up pulling him along.

The Ferris wheel is similar except that its plane of rotation is vertical. The apparent centrifugal force seems periodically to increase and decrease the passenger's weight. When he passes through the bottom of the circle of travel, the centrifugal force appears to push him downward into the seat as if he then weighed more. In reality the seat pushes strongly against him as it keeps him moving in a circle. This force must be strong because it opposes the passenger's weight. At the top of the circle the passenger has the sensation of being somewhat lighter because the apparent centrifugal force is then upward, seemingly pulling him out of his seat. Actually the sensation comes from the fact that the force from the seat is then smaller.

At midpoint of the descent an even stranger sensation is felt. The force from the seat matches the passenger's weight, and the centrifugal force is outward. Hence the passenger feels as though he is about to be thrown forward out of the compartment.

My favorite among the rotating rides is the Rotor, which is a vertical cylinder with a diameter of about 12 feet. The rider stands with his back against the wall as the cylinder begins to spin. When the maximum spin rate is reached, the floor drops away, but the rider remains stuck to the wall. A particularly agile person might be able to squirm enough to get himself into an angled position or even upside down.

Why does the rider stick to the wall? From his perspective a centrifugal force pins him there. The resulting friction between him and the wall prevents him from falling when the floor is removed. A high rate of spin is called for, so that the apparent centrifugal force generates enough friction.

From the perspective of an outside observer the story is different. The rider is constrained to move in a circle because of a force from the wall. This centripetal force is responsible for the friction. Still, the spin rate has to be high if the force from the wall is to generate enough friction to hold the rider in place.

The Rotor at Geauga Lake has a roughly textured wall to increase the friction. With a smoother wall the centripetal force would have to be stronger to keep the rider from slipping. (One would either have to increase the spin rate or build a cylinder with a larger diameter.) Each time I rode the Rotor I was impressed by the overwhelming sensation that a centrifugal force was pushing me against the wall. In reality the wall was pushing on my back.

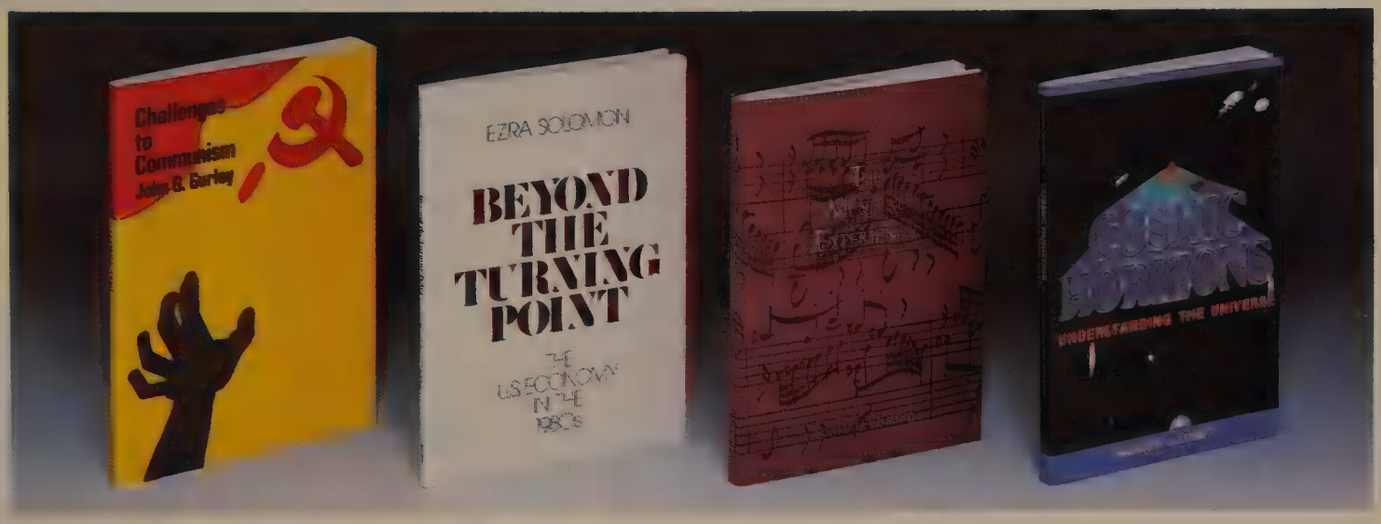
In order for the rider to stay in place his weight (a force vector downward) must not exceed the friction (a force vector upward). The amount of friction can at most be equal to the product of the friction coefficient (which depends on the roughness of the surfaces in contact) and the centripetal force from the wall. I estimated that the spin rate had to be about 30 revolutions per minute to hold me against the wall. Indeed, the apparatus did turn at about that rate.

Several other rides at Geauga Lake involve an apparent centrifugal force. The Muzek Express consists of a series of cars moving on a circular track that traverses several small hills. The diameter of the track is roughly 30 feet. The ride is fast, and so the centrifugal force on a passenger is quite strong. The hills provide extra thrills. Usually two people ride side by side in a car. Since they both feel an outward force, the passenger on the outside is squeezed against the wall of the car by the passenger on the inside. The forces are surprisingly large even if a passenger is small. I cannot avoid being pushed into the wall even when the inside passenger is my young daughter, who weighs less than half what I do.

The Enterprise is a rotating ride with cars individually suspended on radial arms extending from a central hub. As the cars begin to move in a horizontal circle the apparent centrifugal force makes the car rotate outward on the radial arm. Soon the car has rotated almost 90 degrees, and the passenger can see the ground directly below the window that originally was on the inside.

This rotation results from the way the mass of the car and the mass of the passenger are distributed with respect to the suspension axis of the car. A combined centrifugal force operates on the common center of mass of the passenger and

In pursuit of excellence



From communism to the cosmos, economic analysis to music appreciation, these extraordinary new books showcase the work of today's most innovative thinkers. Written by members of Stanford University's renowned faculty, these books, originally published for the exclusive Portable Stanford series, are now available to the public.

Provocative and often controversial, the Portable Stanford has repeatedly been honored for presenting the best of contemporary thought in a highly readable format, designed to be read in an evening or two. You will find incisive thinking, clearly communicated without technical jargon or lengthy footnotes, published for those who are curious about new ideas in fields outside their own.

CHALLENGES TO COMMUNISM

John G. Gurley

Is world communism now in the beginning stages of a long-term decline? This comprehensive study suggests that it is, and explains why. The author also carries forward the theme of his highly acclaimed *Challengers to Capitalism: Marx, Lenin, Stalin, and Mao*, that communism has weakened present-day capitalist countries. Demonstrating how capitalist and communist countries now confront one another in social and economic disarray, Gurley points to the emerging Third World as an increasingly important element in this precarious balance of power.

hardcover: 0-7167-1506-0 $20.00
paperbound: 0-7167-1507-4 $10.95

BEYOND THE TURNING POINT THE U.S. ECONOMY IN THE 1980s

Ezra Solomon

The distinguished economist, a former member of the President's Council of Economic Advisors, examines the economic issues of the '80s both historically and analytically. He traces the source of our problems and explores the outlook for the U.S. in the context of past events.

"Ezra Solomon has brought his high intellect, vast experience, and great insight to this readable and informative analysis of the prospects for our economy. Everyone can learn from reading this book, including those voting on economic policy in Washington."—George P. Shultz, former Secretary of the Treasury

paperbound: 0-7167-1391-8 $7.95

THE MUSICAL EXPERIENCE SOUND, MOVEMENT, ARRIVAL

Leonard G. Ratner

A lively, straightforward exposition of the ways musical elements combine to convey emotion, this is perhaps the ultimate book for those wishing to expand their musical understanding. Using illustrative material from familiar concert music, the author explores musical forms in terms of rhythm, melody, harmony, and texture. The book makes accessible to music lovers the fundamentals of musical style and form. Selected musical examples, recorded on soundsheets bound into the book and playable on any standard phonograph, guide listeners to a fuller appreciation of this appealing art.

hardcover: 0-7167-1475-2 $20.00
paperbound: 0-7167-1476-0 $10.95

COSMIC HORIZONS UNDERSTANDING THE UNIVERSE

Robert Wagoner and Donald Goldsmith

A mind-expanding exploration of the composition of the universe as astronomers now understand it, *Cosmic Horizons* sets in perspective the fundamental principles of modern cosmology. The authors discuss the "big bang" theory, the significance of elementary particles in the development of the universe, and recent advances in particle physics, all of which have led to a more accurate perception of the world around us. This book provides new insights into the sophisticated techniques that allow scientists to probe farther into space—and further back in time.

hardcover: 0-7167-1417-5 $18.95
paperbound: 0-7167-1418-3 $9.95

These books are available at fine bookstores.
To order directly from the publisher, please send a check for the total amount
(plus $1.50 for postage and handling; California, New York, and Utah residents add appropriate sales tax) to:
Order Department, W. H. Freeman and Company, 4419 West 1980 South, Salt Lake City, Utah 84104.



Ford

Created for the State-of-the Artist.



Beethoven was the classic performer of his time. Today, it's the Ford Electronic Premium Sound System. Designed, engineered and quality tested by Ford, specifically for Ford and Lincoln-Mercury automobiles. This highly sophisticated stereo system will faithfully reproduce the musical performance of the most demanding artist. From Beethoven to Bach. From Bach to Rock. And everything in between. The Ford Electronic Premium Sound System. The classic performer in a full line of high performance sound systems by Ford.

Available at your Ford and Lincoln-Mercury dealers today.

Beethoven never heard it so good.

Ford Electrical and Electronics Division

the car. Initially this point lies below the suspension axis. Also acting through the point is the combined weight of the passenger and the car. These two forces compete in orienting the car. Initially gravity pulls the car into the normal orientation, but as the ride moves faster and the centrifugal force gets stronger the car is rotated increasingly out of the vertical.

This much of the ride was disturbing, but the next part almost did me in. Once the ride had reached its highest speed the large arm that held the central hub was turned to make the plane of the moving cars vertical. I was then moving in a vertical circle, being completely upside down at the top and greatly compressed by the forces acting on me at the bottom. I closed my eyes and began to count the prime numbers.

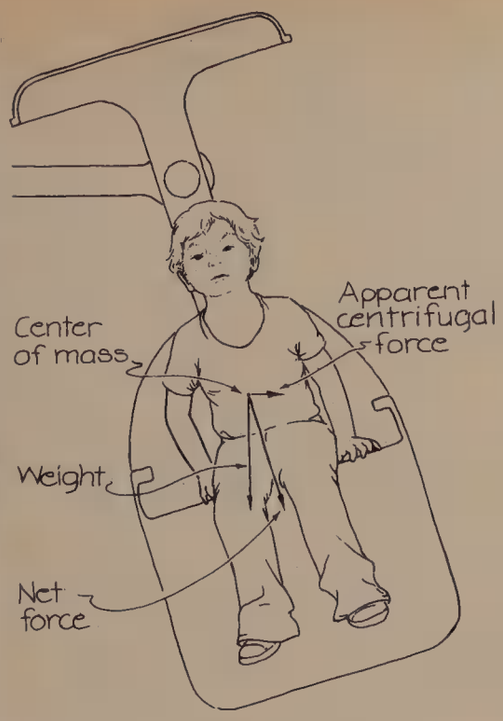
My next ride also held a surprise. It was a set of swings about 20 feet in diameter suspended from a central hub. When the hub began to turn, I moved in a circle below the rim of the hub. As the speed increased, the apparent centrifugal force moved me outward so that I traveled in a larger circle than before. The faster the hub turned, the larger the circle was.

From my perspective three forces affected me. I still had weight, which was directed downward. The chair and its suspension chains provided a second force directed toward the attachment of the chains to the overhead hub. The third force was the fictitious centrifugal force I felt throwing me outward. The angle between the chains and the vertical was set by the balancing of the three forces. When the speed of the ride increased, the angle also increased, so that the forces again balanced.

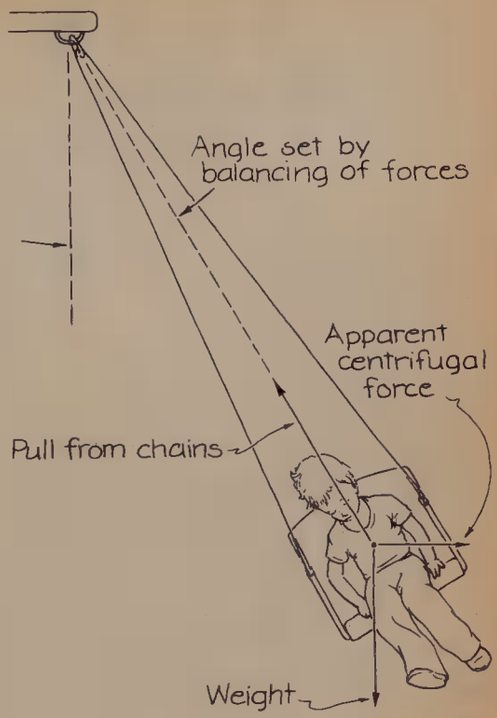
The surprise of the ride was that the hub soon tilted about 10 degrees or so out of the horizontal. Part of my travel around the apparatus was then downhill. My speed increased during the descent as potential energy was converted into kinetic energy. As a result I circled the apparatus in a large radius. In the uphill part of the circle I slowed as the hub was forced to hoist me, and so here I circled with a smaller radius.

I ended my busy day at Geauga Lake with three rides that delivered similar types of motion. The first was the Scrambler, which has long been popular at amusement parks. It consists of a central hub from which several radial arms extend. I call them the primary arms. At the end of each primary arm four secondary arms extend outward. Each one carries at its outer end a car for two or three passengers.

The ride consists of two circular motions. The primary arms rotate steadily about the center of the ride while each set of four secondary arms twirls below the pivot at the end of the primary arm.



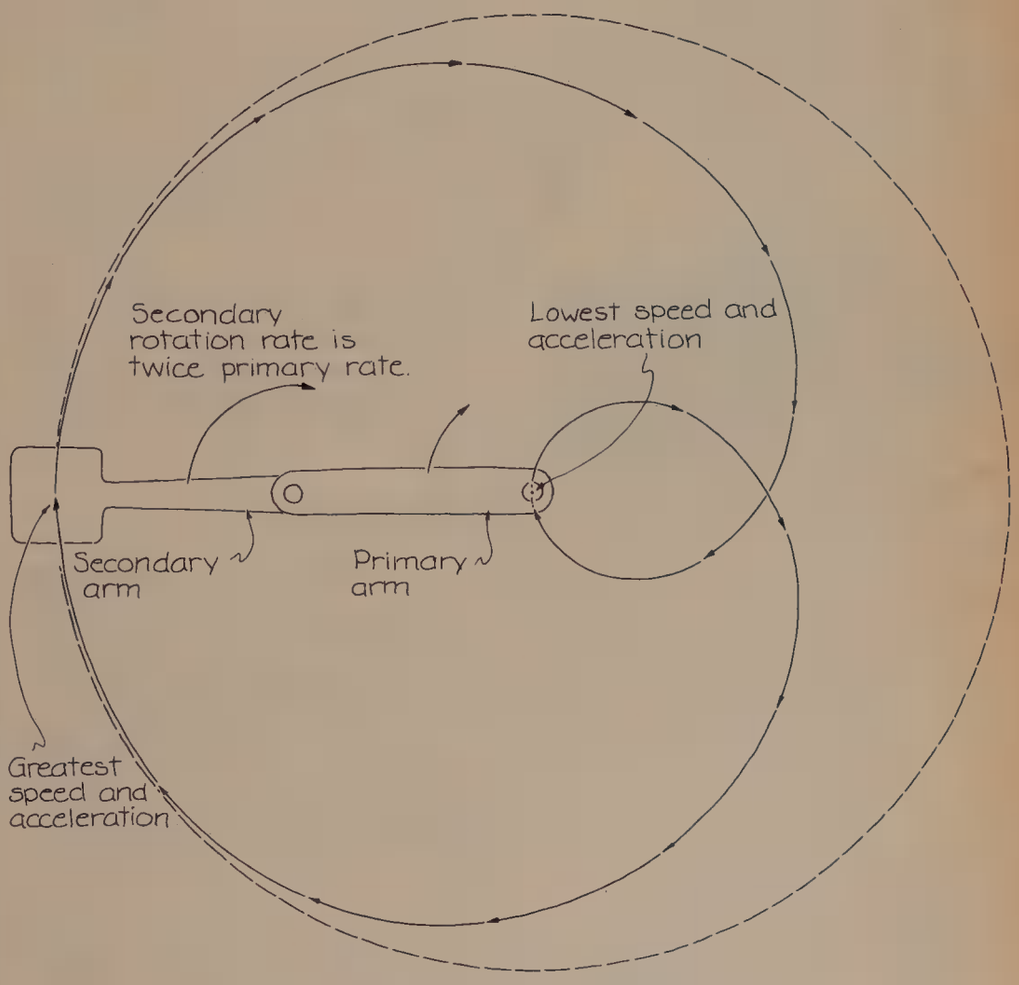
Forces in the Enterprise



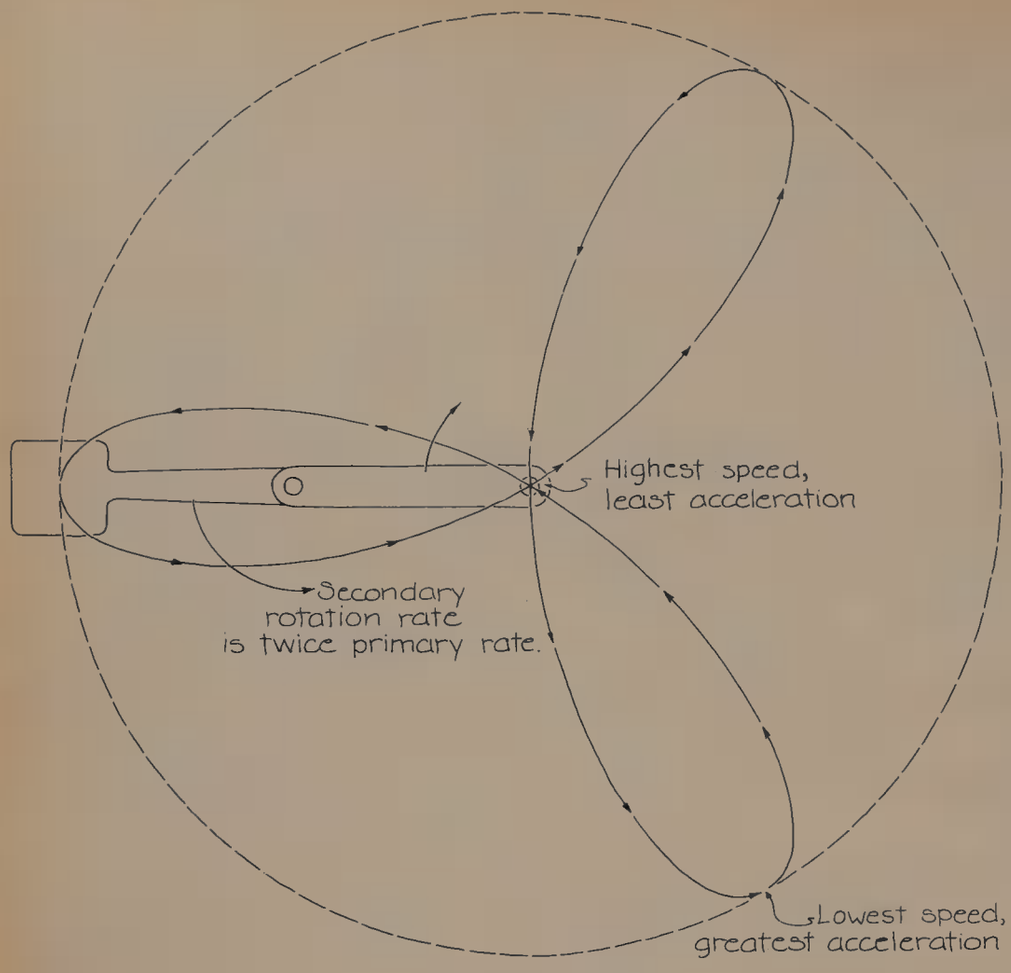
What happens with a rotating swing

From an overhead perspective the primary arms move clockwise, the secondary arms counterclockwise. (In a related ride called Calypso both motions are clockwise.)

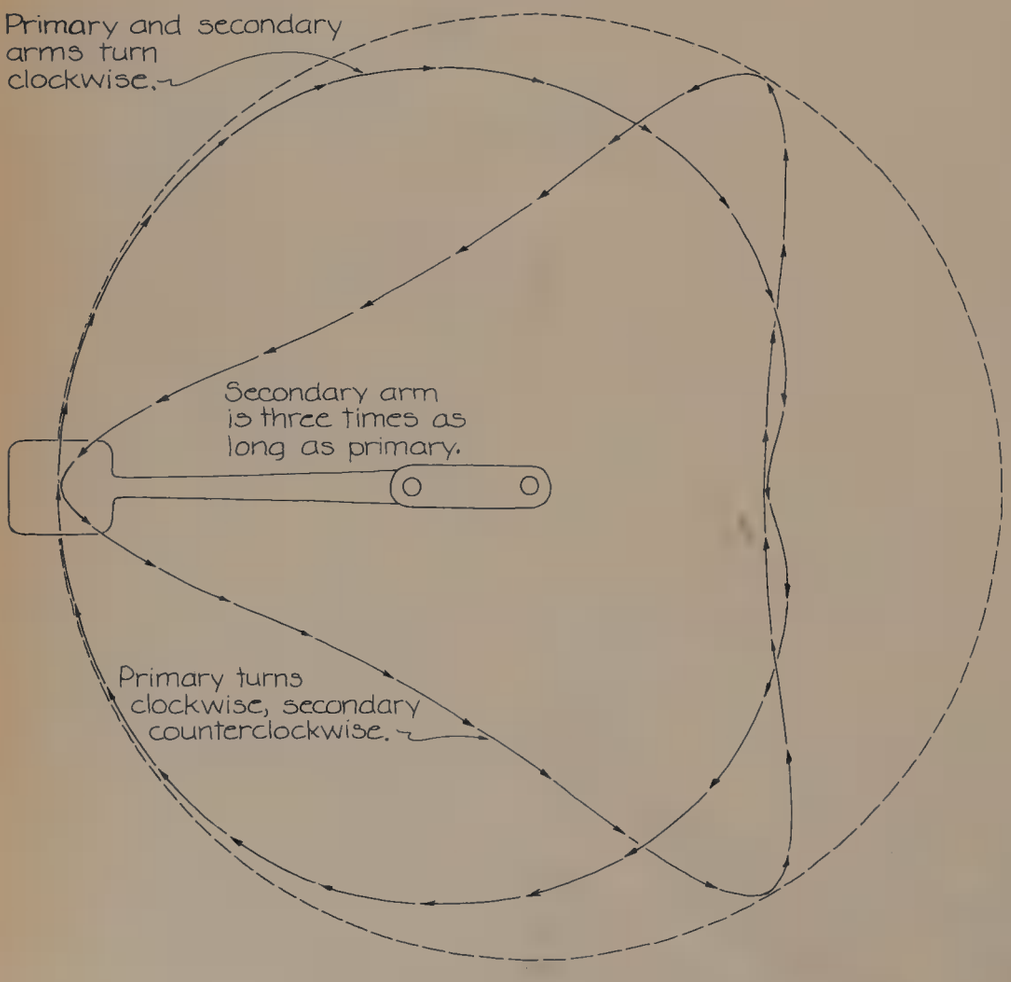
I set about studying the types of motion in rides such as the Scrambler and the Calypso. To model what happens to a rider I focused on a single primary arm (turning clockwise) and a single secondary arm (turning in either direction). As the primary arm makes a full revolution does the passenger loop or spiral? Where are the speed and the acceler-



The motions in a Calypso ride



The motions in a Scrambler ride



The motions in a Calypso or a Scrambler with arms of unequal length

ation greatest? How should the arms rotate to give an unforgettable ride? Should the arms be approximately the same length (as they are in the Scrambler and the Calypso)?

I found that if the arms are the same size and rotate at the same rate and in the same direction, the ride is bound to be rather boring, because the passenger merely goes in a large circle. The ride is not much better if the arms turn in opposite directions. In this arrangement the passenger would travel on a straight line over the center to the opposite side of the ride and then would return on the same line.

A better ride results when the primary and secondary arms rotate at different rates. Suppose the secondary arm turns twice as fast as the primary one. When the primary and secondary arms move in a clockwise direction, as they do in the Calypso, the passenger first spirals in toward the center of the ride and then out again, so that he travels through a loop on the side opposite to the starting point. After spiraling outward he passes through his initial location and begins the trip again.

His speed and acceleration are highest when he is farthest from the center, that is, when he passes through the initial point. They are lowest when he passes over the center of the ride. My calculations approximate the conditions of the Calypso but are off somewhat because to simplify matters I visualized arms of equal length. In order to accommodate all the primary and secondary arms on the Calypso the secondary arms are shorter than the ones in my calculations, so that the cars do not crash near the center of the ride.

If the primary and secondary arms turn in opposite directions, as they do in the Scrambler, a more interesting motion results. At first the passenger moves counterclockwise, but he quickly heads for the center of the ride and then outward again. When the arms are fully extended, he is directed back toward the center. When the primary arm has completed one revolution, the passenger has traveled through a pattern resembling three narrow petals. He is moving at the highest speed when he passes over the center of the ride. Surprisingly, the acceleration there is the lowest. The speed is the lowest and the acceleration the highest when the passenger is farthest from the center.

The low speed at that point results because the circular motion of the secondary arm is carrying the passenger counterclockwise while the motion of the primary arm is clockwise. The two motions oppose each other when the passenger is farthest from the center and augment each other as he is carried close to the center. The high acceleration at the far point develops because

the direction of the velocity out there is changing rapidly.

On my home computer I worked out the paths for other conditions. If the primary arm is much longer than the secondary arm, the passenger might spiral toward and then away from the center. In other cases he would move through a path consisting of a series of cusps or loops superposed on a large circle. Another interesting situation arises when the secondary arm is longer than the primary one. If the secondary arm turns slower than the primary one and in the same direction, the passenger will spiral gradually toward the center and then away from it. If the motions are in opposite directions, the ride will include an abrupt change in direction.

My last ride of the day was on the Tilt-A-Whirl. I sat in a compartment that was free to move around a small circular track, pivoting on a center point at my feet. Six of these compartments, running on a hilly track, moved around the center of the apparatus while they were also pivoting on their individual center points. The result was three types of motion. One was a basic counterclockwise circling about the center of the entire ride. The second was a smaller circling of the compartment in either direction. The third was the vertical motion over the hills.

The interesting part was that I could often control the small circling of my compartment by anticipating the hills and shifting my weight. When the compartment was turning in its small circle and beginning to descend from a hill in the larger motion, I threw my weight in the direction of the compartment's turn. I was transforming some of the potential energy of my body (from being up on the hill) into kinetic energy applied to the rotation of the compartment.

When I timed this exercise correctly, I set the compartment spinning rapidly. My experience was similar to that in the preceding two rides. If the spin direction was in the same direction as the large-scale circling of the ride, the speed and centrifugal force were quite large when I was farthest from the center of the ride. If the spin was in the opposite direction, the acceleration was high when I was far from the center but the speed was high closer to the center.

You might explore the amusement parks near you for other rides. One that I have heard described, but lack the courage to even look at, is Demon Drop. The victim—sorry, the passenger—is secured in a chair that is lifted 131 feet and then dropped in a virtually free fall to the ground. The huge kinetic energy of the ride is apparently dissipated when the apparatus curves into a horizontal section of track at the bottom of the fall. I have no intention whatever of verifying this assumption.

WHAT'S BETTER THAN SPEED READING?

SPEED LEARNING

(SPEED PLUS COMPREHENSION)

Speed Learning is replacing speed reading. It's easy to learn . . . lasts a lifetime . . . applies to everything you read . . . and is the only accredited course with special editions for professional reading.

Do you have too much to read and too little time to read it? Do you mentally pronounce each word as you read? Do you frequently have to go back and re-read words or whole paragraphs you just finished reading? Do you have trouble concentrating? Do you quickly forget most of what you read?

If you answer "yes" to any of these questions—then here at last is the practical help you've been waiting for. Whether you read for business or pleasure, school or college, you will build exceptional skills from this major breakthrough in effective reading, created by Dr. Russell Stauffer at the University of Delaware.

Not just "speed reading"—but speed reading-thinking-understanding-remembering-and-learning

The new *Speed Learning Program* shows you step-by-proven-step how to increase your reading skill and speed, so you understand more, remember more and use more of everything you read.

Imagine the new freedom you'll have when you learn how to dash through all types of reading material *at least* twice as fast as you do now, and with greater comprehension. Think of being able to get on top of the avalanche of newspapers, magazines and correspondence you have to read . . . finishing a stimulating book and retaining facts and details more clearly and with greater accuracy than ever before.

What makes Speed Learning so successful?

In just a few spare minutes a day of easy reading and exciting listening, you discover an entirely new way to read and think—a radical departure from anything you have ever seen or heard about. Research shows that reading is 95% *thinking* and only 5% eye movement. Yet most of today's speed reading programs spend their time teaching you rapid eye movement (5% of the problem) and ignore the most important part (95%) *thinking*. In brief, *Speed Learning* gives you what speed reading *can't*.

Executives, students, professional people, men and women in all walks of life from 15 to 70 have benefited from this program. Speed Learning is a fully accredited course . . . costing only ⅕ the price of less effective speed reading classroom courses. Now you can examine the same easy, practical and proven methods at home . . . in spare time . . . without risking a penny.

Examine Speed Learning FREE for 15 days

You will be thrilled at how quickly this program will begin to develop new thinking and reading skills. After listening to just one cassette and reading the preface you will quickly see how you can achieve increases in both the speed at which you read and in the amount you understand and remember.

You must be delighted with what you see or you pay nothing. Examine this remarkable program for 15 days. If, at the end of that time you are not convinced that you would like to master Speed Learning, simply return the program and owe nothing.



learn INCORPORATED Dept. DZ-23, 113 Gaither Drive, Mount Laurel, NJ 08054

YES! Please send the materials checked below:

☐ Standard Speed Learning program @ $99.95 plus $5 postage and handling

☐ Special Edition Speed Learning program @ $125 plus $5 postage and handling (check one below):

☐ Medical ☐ Data Processing ☐ Science/Engineering ☐ Management ☐ Finance/Accounting

Check the method of payment below: NJ Residents add 6% sales tax

☐ Check or money order enclosed

☐ Charge my credit card:

☐ Visa ☐ Master Card ☐ American Express

Card #_____ Exp. Date_____

I understand that I may examine the materials for 15 days. At the end of that time, if I am not delighted in every way I may return them for a full refund. No questions asked.

Name_____

Address_____

City_____ State_____ Zip_____

Signature_____

If you don't already own a cassette player, you may order this Deluxe Cassette Recorder for only $49.95. (Includes handling and delivery.)

Check here to order ☐



Outside USA add $10 per item—Airmail extra

BIBLIOGRAPHY

Readers interested in further explanation of the subjects covered by the articles in this issue may find the following lists of publications helpful.

COMPUTER RECREATIONS

AUTOMATA STUDIES. Edited by Claude E. Shannon and John McCarthy. Princeton University Press, 1956.

MATHEMATICAL GAMES. Martin Gardner in *Scientific American,* Vol. 223, No. 4, pages 120–123, October, 1970, and Vol. 224, No. 2, pages 112–117, February, 1971.

WINNING WAYS FOR YOUR MATHEMATICAL PLAYS. Elwyn R. Berlekamp, John H. Conway and Richard K. Guy. Academic Press, 1982.

LOTUS 1-2-3 USER'S MANUAL. John Posner, Steven E. Miller, Ezra Gottheil and Jeff Hill. Lotus Development Corporation, 1983.

LABOR-INTENSIVE AGRICULTURE

HIRED HANDS: SEASONAL FARM WORKERS IN THE UNITED STATES. Stephen H. Sosnick. McNally & Loftin, 1978.

THE H-2 PROGRAM AND NONIMMIGRANTS. Hearing before the Subcommittee on Immigration and Refugee Policy, November 30, 1981 (J-97-85). U.S. Government Printing Office, 1981.

THE DEMAND FOR SEASONAL FARM LABOR IN CALIFORNIA. Philip L. Martin and Richard Mines. Working Paper 83–9, University of California at Davis, 1983.

SEASONAL FARM LABOR IN THE UNITED STATES. Edited by Robert Emerson. Iowa State University Press, 1983.

THE ENGINEERING OF MAGNETIC FUSION REACTORS

RECENT DEVELOPMENTS IN THE DESIGN OF CONCEPTUAL FUSION REACTORS. F. L. Ribe in *Nuclear Technology,* Vol. 34, No. 2, pages 179–208; July, 1977.

PROGRESS TOWARD A TOKAMAK FUSION REACTOR. Harold P. Furth in *Scientific American,* Vol. 241, No. 2, pages 50–61; August, 1979.

MAGNETIC FUSION REACTORS. Robert W. Conn in *Fusion,* edited by Edward Teller. Academic Press, 1981.

THE LARGE-SCALE STRUCTURE OF THE UNIVERSE

THE FIRST THREE MINUTES: A MODERN VIEW OF THE ORIGIN OF THE UNIVERSE. Steven Weinberg. Basic Books, Inc., 1977.

THE BIG BANG: THE CREATION AND EVOLUTION OF THE UNIVERSE. Joseph Silk. W. H. Freeman and Company, 1980.

THE GROWTH OF STRUCTURE IN THE UNIVERSE. J. R. Gott in *Physical Cosmology,* edited by Roger Balian, Jean Adouze and David N. Schramm. Elsevier-North Holland, Inc., 1980.

THE LARGE-SCALE STRUCTURE OF THE UNIVERSE. P. J. E. Peebles. Princeton University Press, 1980.

MODERN COSMOLOGY. Ya. B. Zel'dovich in *Highlights of Astronomy, Vol. 6,* edited by R. West. D. Reidel Publishing Co., 1983.

THE PROCESSING OF RNA

RIBONUCLEIC ACIDS FROM ANIMAL CELLS. James E. Darnell, Jr., in *Bacteriological Reviews,* Vol. 32, No. 3, pages 262–290; September, 1968.

THE DEFINITION OF A LARGE VIRAL TRANSCRIPTION UNIT LATE IN AD2 INFECTION OF HELA CELLS: MAPPING OF NASCENT RNA MOLECULES LABELED IN ISOLATED NUCLEI. Jeffrey Weber, Warren Jelinek and James E. Darnell, Jr., in *Cell,* Vol. 10, No. 4, pages 611–616; April, 1977.

SPLICED SEGMENTS AT THE 5′ TERMINUS OF ADENOVIRUS 2 LATE MRNA. Susan M. Berget, Claire Moore and Philip A. Sharp in *Proceedings of the National Academy of Sciences of the United States of America,* Vol. 74, No. 8, pages 3171–3175; August, 1977.

AN AMAZING SEQUENCE ARRANGEMENT AT THE 5′ ENDS OF ADENOVIRUS 2 MESSENGER RNA. Louise T. Chow, Richard E. Gelinas, Thomas R. Broker and Richard J. Roberts in *Cell,* Vol. 12, No. 1, pages 1–8; September, 1977.

THE INITIATION SITES FOR RNA TRANSCRIPTION IN AD2 DNA. Ronald M. Evans, Nigel Fraser, Edward Ziff, Jeffrey Weber and James E. Darnell in *Cell,* Vol. 12, No. 3, pages 733–739; November, 1977.

IMPLICATIONS OF RNA · RNA SPLICING IN EVOLUTION OF EUKARYOTIC CELLS. James E. Darnell, Jr., in *Science,* Vol. 202, No. 4374, pages 1257–1260; December 22, 1978.

VARIETY IN THE LEVEL OF GENE CONTROL IN EUKARYOTIC CELLS. James E. Darnell, Jr., in *Nature,* Vol. 297, No. 5865, pages 365–371; June 3, 1982.

THE DEAD SEA

SOME GEOCHEMICAL ASPECTS OF THE DEAD SEA AND THE QUESTION OF ITS AGE. Yaacov K. Bentor in *Geochimica et Cosmochimica Acta,* Vol. 25, No. 4, pages 239–260; 1961.

THE MICROBIOLOGY AND BIOGEOCHEMISTRY OF THE DEAD SEA. A. Nissenbaum in *Microbial Ecology,* Vol. 2, pages 139–151; 1975.

THE DEAD SEA: DEEPENING OF THE MIXOLIMNION SIGNIFIES THE OVERTURE TO OVERTURN OF THE WATER COLUMN. I. Steinhorn, G. Assaf, J. R. Gat, A. Nishry, A. Nissenbaum, M. Stiller, M. Beyth, D. Neev, R. Garber, G. M. Friedman and W. Weiss in *Science,* Vol. 206, No. 4414, pages 55–57; October 5, 1977.

SIX MILLENNIUMS OF BUFFALO KILLS

AN INTRODUCTION TO THE ECOLOGY OF EARLY PREHISTORIC COMMUNAL BISON HUNTING AMONG NORTHERN PLAINS INDIANS. George W. Arthur. National Museum of Man Mercury Series, Archaeological Survey of Canada, Paper 40, 1975.

BISON PROCUREMENT AND UTILIZATION: A SYMPOSIUM. Edited by Leslie B. Davis and Michael Wilson. Memoir 14, Plains Anthropologist, 1978.

PREHISTORIC HUNTERS OF THE HIGH PLAINS. George C. Frison. Academic Press, 1978.

THE EXTINCTION OF THE AMMONITES

GEOMETRIC ANALYSIS OF SHELL COILING: COILING IN AMMONOIDS. David M. Raup in *Journal of Paleontology,* Vol. 41, No. 1, pages 43–65; January, 1967.

HYDROMECHANICAL DESIGN OF FOSSIL CEPHALOPODS. J. A. Chamberlain in *Ammonoidea: Evolution, Classification, Mode of Life, and Geological Usefulness of a Major Fossil Group,* edited by M. R. House and J. R. Senior. Academic Press, 1980.

CHAMBER REFILLING IN *NAUTILUS.* P. Ward and L. Greenwald in *Journal of the Marine Biological Association of the U.K.,* Vol. 62, pages 469–475; 1981.

SHELL SCULPTURE AS A DEFENSIVE ADAPTATION IN AMMONOIDS. P. Ward in *Paleobiology,* Vol. 7, No. 1, pages 96–100; 1981.

BILATERAL NEGOTIATIONS AND THE ARMS RACE

PROGRESS IN ARMS CONTROL? Introductions by Bruce M. Russett and Bruce G. Blair. W. H. Freeman and Company, 1979.

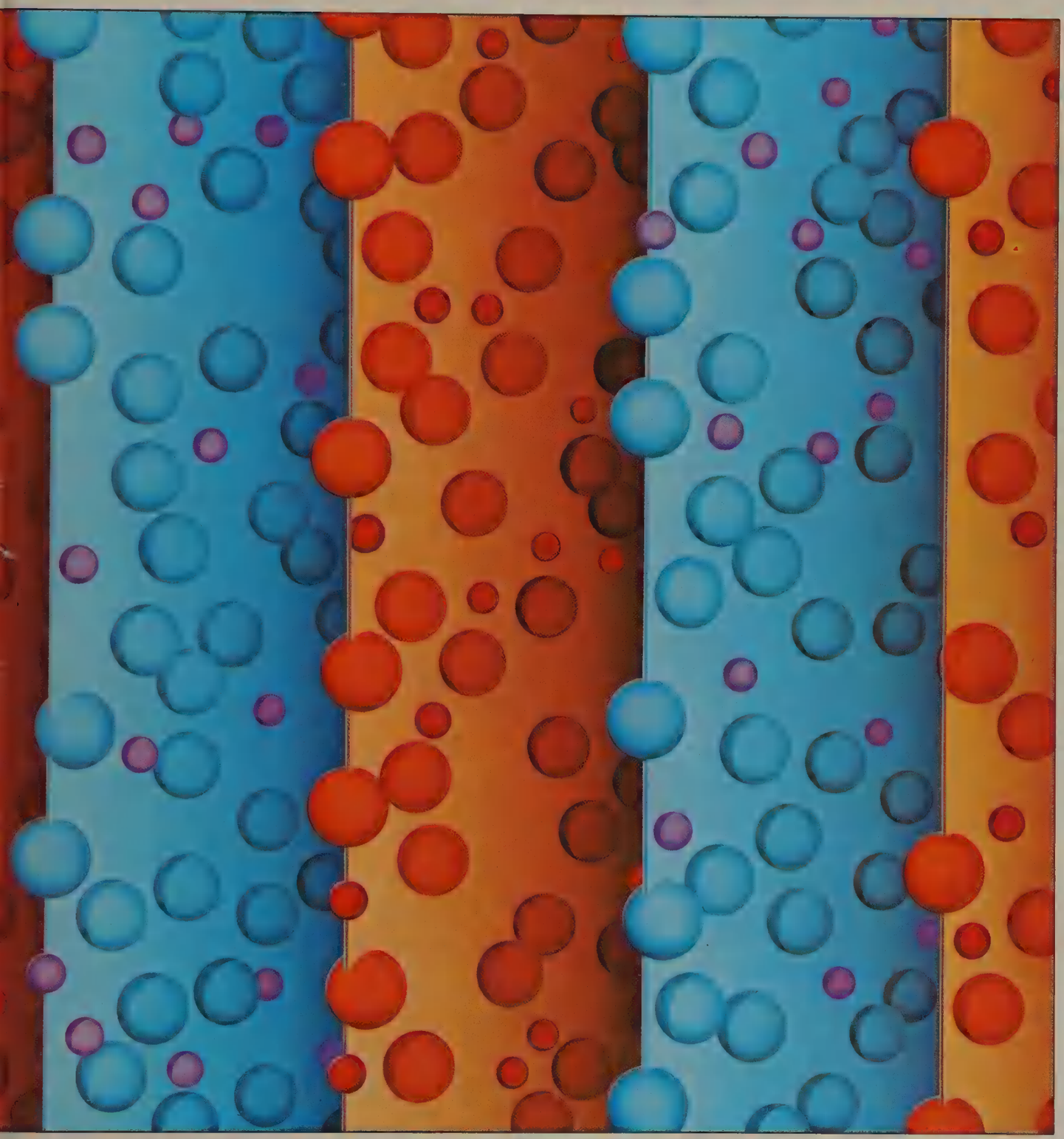
ARMS CONTROL AND DISARMAMENT AGREEMENTS. United States Arms Control and Disarmament Agency, 1982.

THE AMATEUR SCIENTIST

HARRY G. TRAVER: LEGENDS OF TERROR. Richard Munch. Amusement Park Books, Inc., Mentor, Ohio, 1982.

Marygrove College Library
Detroit, Michigan 48221
PLEASE DO NOT REMOVE

SCIENTIFIC AMERICAN



SEMICONDUCTOR SUPERLATTICES

$2.50

November 1983



INTRODUCING RENAULT ENCORE

Let go, for a minute, of all your preconceptions about automotive styling and design. Allow your eyes to study Renault's new hatchback achievement: Encore. Created with a distinctive European look, and a European outlook. A panoramic hatchback that opens to fit your lifestyle. A 60/40 split rear seat configuration that provides a fitting and functional access to cargo. Instrumentation that reads clearly, and is clearly accessible.

*Use for comparison. Your mileage may vary with speed, trip length, weather. Actual highway and CA figures lower. **Manufacturer's suggested retail price. Tax, license, destination charges, rear deck spoiler $72, wheel covers $90, and regional equipment extra. **Built by American Motors**

Established 1845 **SCIENTIFIC AMERICAN** November 1983 Volume 249 Number 5

ARTICLES

38 **THE UNCERTAINTIES OF A PREEMPTIVE NUCLEAR ATTACK, by Matthew Bunn and Kosta Tsipis** They suggest that the vulnerability of U.S. land-based missiles is overstated.

48 **SLURRY PIPELINES, by Edward J. Wasp**
Ore concentrates, coal and other commodities are transported long distances suspended in water.

56 **CENTAURUS A: THE NEAREST ACTIVE GALAXY, by Jack O. Burns and R. Marcus Price** Active galaxies radiate a million times more energy than others. How do they do it?

126 **A MOLECULAR BASIS OF CANCER, by Robert A. Weinberg**
With one oncogene a point mutation changes just one amino acid in the protein the gene encodes.

144 **SOLID-STATE SUPERLATTICES, by Gottfried H. Döhler**
Crystals made up of semiconductors in thin layers hold much interest for physics and technology.

152 **THE SOCIAL ARCHAEOLOGY OF MEGALITHIC MONUMENTS, by Colin Renfrew**
The henges of Neolithic Europe are associated with the rise of centralized political control.

164 **A SEED-EATING BEETLE'S ADAPTATIONS TO A POISONOUS SEED, by Gerald A. Rosenthal** The beetle larvae not only circumvent a toxic amino acid but also utilize it.

172 **THE ERUPTION OF KRAKATAU, by Peter Francis and Stephen Self**
Exactly what happened in the Sunda Straits 100 years ago is only now beginning to be understood.

DEPARTMENTS

6 LETTERS
11 50 AND 100 YEARS AGO
12 THE AUTHORS
18 COMPUTER RECREATIONS
32 BOOKS
70 SCIENCE AND THE CITIZEN
188 THE AMATEUR SCIENTIST
198 BIBLIOGRAPHY

BOARD OF EDITORS Gerard Piel (Publisher), Dennis Flanagan (Editor), Brian P. Hayes (Associate Editor), Philip Morrison (Book Editor), John M. Benditt, Peter G. Brown, Michael Feirtag, Robert Kunzig, Diana Lutz, Jonathan B. Piel, John Purcell, James T. Rogers, Armand Schwab, Jr., Joseph Wisnovsky

ART DEPARTMENT Samuel L. Howard (Art Director), Steven R. Black (Assistant Art Director), Ilil Arbel, Edward Bell

PRODUCTION DEPARTMENT Richard Sasso (Production Manager), Carol Eisler and Leo J. Petruzzi (Assistants to the Production Manager), Carol Hansen (Electronic Composition Manager), Carol Albert, Karen Friedman, Lori Mogol, Karen O'Connor, Julio E. Xavier

COPY DEPARTMENT Sally Porter Jenks (Copy Chief), Debra Q. Bennett, Mary Knight, Dorothy R. Patterson

GENERAL MANAGER George S. Conn

ADVERTISING DIRECTOR C. John Kirby

CIRCULATION MANAGER William H. Yokel

SECRETARY Arlene Wright

SCIENTIFIC AMERICAN (ISSN 0036-8733), PUBLISHED MONTHLY BY SCIENTIFIC AMERICAN, INC., 415 MADISON AVENUE, NEW YORK, N.Y. 10017. COPYRIGHT © 1983 BY SCIENTIFIC AMERICAN, INC. ALL RIGHTS RESERVED. PRINTED IN THE U.S.A. NO PART OF THIS ISSUE MAY BE REPRODUCED BY ANY MECHANICAL, PHOTOGRAPHIC OR ELECTRONIC PROCESS, OR IN THE FORM OF A PHONOGRAPHIC RECORDING, NOR MAY IT BE STORED IN A RETRIEVAL SYSTEM, TRANSMITTED OR OTHERWISE COPIED FOR PUBLIC OR PRIVATE USE WITHOUT WRITTEN PERMISSION OF THE PUBLISHER. SECOND-CLASS POSTAGE PAID AT NEW YORK, N.Y., AND AT ADDITIONAL MAILING OFFICES. AUTHORIZED AS SECOND-CLASS MAIL BY THE POST OFFICE DEPARTMENT, OTTAWA, CANADA, AND FOR PAYMENT OF POSTAGE IN CASH. SUBSCRIPTION RATE: $24 PER YEAR IN THE U.S. AND ITS POSSESSIONS; $33 PER YEAR IN ALL OTHER COUNTRIES. POSTMASTER: SEND ADDRESS CHANGES TO SCIENTIFIC AMERICAN, 415 MADISON AVENUE, NEW YORK, N.Y. 10017.

SCIENTIFIC AMERICAN

CORRESPONDENCE

Offprints of more than 1,000 selected articles from earlier issues of this magazine, listed in an annual catalogue, are available at 75 cents each. Correspondence, orders and requests for the catalogue should be addressed to W. H. Freeman and Company, 4419 West 1980 South, Salt Lake City, UT 84121. Offprints adopted for classroom use may be ordered direct or through a college bookstore. Sets of 10 or more Offprints are collated by the publisher and are delivered as sets to bookstores.

Photocopying rights are hereby granted by Scientific American, Inc., to libraries and others registered with the Copyright Clearance Center (CCC) to photocopy articles in this issue of SCIENTIFIC AMERICAN for the flat fee of $2 per copy of each article or any part thereof. Such clearance does not extend to the photocopying of articles for promotion or other commercial purposes. Correspondence and payment should be addressed to Copyright Clearance Center, Inc., 21 Congress Street, Salem, MA 01970. Specify CCC Reference Number ISSN 0036–8733/83. $2.00 + 0.00.

Editorial correspondence should be addressed to The Editors, SCIENTIFIC AMERICAN, 415 Madison Avenue, New York, NY 10017. Manuscripts are submitted at the authors' risk and will not be returned unless they are accompanied by postage.

Advertising correspondence should be addressed to C. John Kirby, Advertising Director, SCIENTIFIC AMERICAN, 415 Madison Avenue, New York, NY 10017.

Subscription correspondence should be addressed to Subscription Manager, SCIENTIFIC AMERICAN, P.O. Box 5969, New York, NY 10017. The date of the last issue on your subscription is shown in the upper right-hand corner of each month's mailing label. For change of address notify us at least four weeks in advance. Please send your old address (if convenient, on a mailing label of a recent issue) as well as the new one.

Name

New Address

Street

City

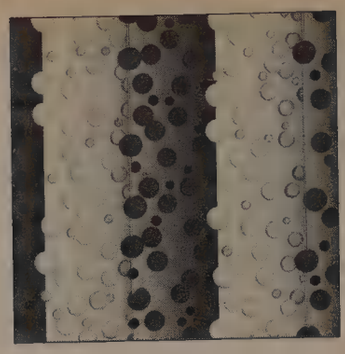
State and ZIP

Old Address

Street

City

State and ZIP



THE COVER

The painting on the cover represents a superlattice: a crystal whose remarkable electronic and optical properties arise from ultrathin layers of semiconductor (see "Solid-State Superlattices," by Gottfried H. Döhler, page 144). It is a doping superlattice, which consists of a single semiconductor in layers "doped" with different impurities. The orange planes are *n* layers: they are doped with donor atoms (*orange spheres*), which give up electrons and thus become positive ions. The blue planes are *p* layers: they are doped with acceptor atoms (*blue spheres*), which bind electrons and thus become negative ions. A voltage applied to the crystal sends free electrons (*small red spheres*) into the *n* layers, and "holes," or the absence of electrons (*small purple spheres*), into the *p* layers. In a typical semiconductor these charge carriers recombine in billionths of a second, and the energy of each electron-hole pair is dissipated by a photon (a quantum of light). In a superlattice the carriers can survive for several hours.

THE ILLUSTRATIONS

Cover painting by George V. Kelvin

Page	Source
18–21	Edward Bell
24D	Ilil Arbel (*top*), Edward Bell (*bottom*)
25–28	Edward Bell
39	U.S. Air Force
40–46	George V. Kelvin
47	Ilil Arbel
49	Edward J. Wasp
50–55	Walken Graphics
57	Jack O. Burns, Eric D. Feigelson and Ethan J. Schreier
58	Jack O. Burns and R. Marcus Price, University of New Mexico
60	Ian Worpole
61	Jack O. Burns, Eric D. Feigelson and Ethan J. Schreier
62–63	Ian Worpole
64	Jack O. Burns, Eric D. Feigelson and Ethan J. Schreier
65	Ian Worpole
127	Erika A. Hartwieg and Jonathan A. King, Massachusetts Institute of Technology
127A–139	Bunji Tagawa
140	Ravi Dhar
142	Bunji Tagawa
145	Pierre M. Petroff and Arthur C. Gossard, Bell Laboratories
146–151	Allen Beechel
153	Colin Renfrew, University of Cambridge
154	Andrew Tomko
157	West Air Photography
158–160	Andrew Tomko
161	Patricia J. Wynne after Alec Daykin
162	Colin Renfrew, University of Cambridge
165	James E. Mattler
166–171	Patricia J. Wynne
173	*Album of Krakatau*
174–183	Todd Pink
186	Maurice Krafft
189	Jearl Walker
190–196	Michael Goodman

Smith-Corona brings portable typewriters into the Computer Age.



Now Smith-Corona® takes computer technology and puts it into a new electronic *portable* typewriter. Typing has never been so simple!

The Smith-Corona electronic portable typewriter has features never even possible without today's advanced technology.

Simple One-Touch Memory Correction.

The Smith-Corona electronic portable has a memory, so it remembers what you type. Then, when you press the correct key, a letter, a word, even a line is lifted right off the page.

Simple Automatic Operations.

No more time-consuming operations like counting and backspacing. Just press the appropriate code key and type. Automatic centering, underlining, decimal tabulation, relocation, carrier return, flush right and reverse tabulation! The Smith-Corona electronic portable produces perfect results every time.

Simple Computer Compatibility.

The Smith-Corona electronic portable doubles as a computer printer. Simply connect the typewriter to the computer with optional Messenger Module.

Impressive Print Quality.

Add to these features a print quality as impressive as on big, expensive office machines. And the Smith-Corona electronic portable can be adjusted to type 10, 12 or 15 spaces per inch. Inexpensive ribbons, and printwheels which offer a choice of type styles, couldn't be easier to change.

Impressive Name.

Over the years, Smith-Corona has always been the leader in portable typewriter technology. No wonder our name has set the standards of the industry!

Impressive Price.

Best of all, Smith-Corona electronic portables are truly affordable. So you can have advanced office typing technology at home.

Try a new Smith-Corona electronic portable at your dealer's. And see for yourself how marvelous Computer Age typing can be.

Some electronic features not available on all models.

Smith-Corona



LEARN TO SPEAK
SPANISH
WHILE DRIVING TO WORK
WITH THE AMR® PATENTED
BI-LINGUAL
STEREO-SELFTEST™ METHOD

YOU CAN NOT HELP BUT PICK UP
A SECOND LANGUAGE
WITHOUT OPENING A BOOK

SPANISH-FRENCH-GERMAN-ITALIAN
CHINESE-JAPANESE-ENGLISH

Call us at AMR and a staff member will gladly answer your questions

☎ 1-800-323-2500 ☎
IL Residents 312-858-1900

AMR®
800 Roosevelt Rd.
Glen Ellyn, Illinois
60137
EDUCATIONAL SYSTEMS

"On Nov. 17th, adopt a friend who smokes."



Help a friend get through the day without a cigarette. They might just quit forever. And that's important. Because good friends are hard to find. And even tougher to lose.

THE GREAT AMERICAN SMOKEOUT

AMERICAN CANCER SOCIETY®

LETTERS

Sirs:

Philip Morrison writes in his review of Kenneth J. Hsü's book "The Mediterranean Was a Desert: A Voyage of the *Glomar Challenger*" [SCIENTIFIC AMERICAN, September] that the realization that the bed of the Mediterranean was once a desert was a "bright flash of inference... in the lofty tradition of field geology [going] back to Louis Agassiz and James Hutton." Actually Hutton imagined this possibility almost two centuries before it was confirmed. I quote from Hutton's *Theory of the Earth,* Volume 1, page 75, although he said the same thing in a paper of 1788:

"The formation of salt at the bottom of the sea, without the assistance of subterranean fire, is not a thing unsupposable, as at first sight it might appear. Let us but suppose a rock placed across the gut of Gibraltar (a case nowise unnatural), and the bottom of the Mediterranean would be certainly filled with salt, because the evaporation from the surface of that sea exceeds the measure of its supply."

KEITH J. TINKLER

Brock University
St. Catharines, Ont.

Sirs:

"The Stave Churches of Norway," by Petter Aune, Ronald L. Sack and Arne Selberg [SCIENTIFIC AMERICAN, August], commented on the assimilation of Norse elements in the architecture and decoration of these 10th-century Christian structures. The diagrams and photographs enhanced the text in a beautifully woven manner. Sack's photograph of an exterior carving was not, however, properly described. The designs are an intermingling not of Norse and Christian elements but of diverse cultural influences. Besides the Christian inspiration in the plant tendrils, the dragon was introduced by Roman legions and became the guardian of the grave; the griffin was originally found in Hittite symbolism. The dragon and griffin are not intrinsic to Norse mythology but are cultural accretions of a later period.

WILLIAM ROBA

Scott Community College
Bettendorf, Iowa

Sirs:

"Chapman's rule" for judging a fly ball ["Science and the Citizen," SCIENTIFIC AMERICAN, April] would limit successful outfielders to college graduates with a good grade in mathematics. Who else could judge whether the tangent of the ball's angle of elevation rose at a constant rate?

May I, as an expert on the subject ever since I was a foreign student and was given a free ticket to a game between the Pittsburgh Pirates and the New York Mets, offer a simpler interpretation? The correct position of the outfielder is to stand where the ball appears to rise above the head of the batter at a constant speed. Hoping to have served the interest of the great game, play ball!

GUSTAV WINKLER

Flensburg, Federal Republic
of Germany

Sirs:

Michel Bur has done an admirable job in assessing the historical significance of the motte ["The Social Influence of the Motte-and-Bailey Castle," by Michel Bur; SCIENTIFIC AMERICAN, May]. Bur reports that the network of the first French mottes had no connection with existing population centers. I suspect that engineering factors may have played a significant role in 10th-century motte siting. Consider a motte 100 meters in diameter at its base and 20 meters high. Such an earthwork contains 410,000 cubic meters of earth. This is a substantial amount of material to excavate, transport, dump and form. Even with modern earth-moving equipment such a construction job, although it is simple, is not easy or quick. For example, if one assumes that the excavation is done in earth with a Euclid loader having a capacity of 790 cubic meters per hour and that the earth is hauled over an average distance of one kilometer with no more than a 3 percent grade by seven tractor-pulled bottom dump wagons with a capacity of 18 cubic meters working 10 hours per day, 112 days would be needed to complete the job.

The 10th-century engineers would therefore have faced a significant challenge. If only manually pushed wheelbarrows were used, I estimate that 56 million loads of 140 kilograms each would have been required. Larger loads would of course have been possible with domestic animals.

It seems reasonable that the determination of motte location included the consideration of factors such as the availability of earth, favorable topography and climate, suitable indigenous labor and the ability to protect the construction site for the many months, and possibly years, needed to complete the fortifications.

ROBERT E. SCHAFRIK, PH.D. P.E.

Dayton, Ohio



Test Drive the Leica R4.

We invite you to experience the perfection of the Leica® R4. We invite you to hold, inspect, and then test the Leica that has been acclaimed "...a miracle of modern technology." A limited production SLR of astonishing versatility, precision and beauty. With flawless Leitz® optics, two exposure measurement systems, five operation modes, and a series of advances unequalled in the history of photography.

The Leica R4 is Leica's proudest achievement. A photographic instrument designed for perfection, not price. Like every Leica since the beginning, it is lavished with attention to detail. Hand finishing. Hand assembly. Hand testing. No shortcuts or compromises. Because the Leica R4 is designed and built never to fail. And to have no limit to its useful life.

We invite you to experience the Leica R4. And through it enter Leica's world of superior photography. To see and appreciate for yourself the perfection of the photographic image that is uniquely Leica's.

You can test drive the Leica R4 now, at your local Leica dealer. And to show his appreciation for your visit and your interest, your Leica dealer will enter your name in a special drawing for the ultimate sports coupe—the BMW 633CSi.* We think you'll find this high performance German import, the perfect accessory to your new Leica.

For the name of your nearest Leica dealer write E. Leitz, Inc., 24 Link Drive, Rockleigh, N.J. 07647 or call (800) 223-0507, ext. 801. In New Jersey call (800) 223-0509, ext. 801.

And you could win a BMW 633 CSi to drive it home in.

*No purchase necessary. Entries must be made through your Leica dealer before January 31, 1984. The winner will be notified by March 30, 1984. Void where prohibited by law.

The renaissance man is alive and well.

This once endangered species has made an astounding comeback.

Take Tom McWilliams.

He grew up with lots of skis, bicycles and microscopes.

And an Apple® II Personal Computer.

Using the latter to design a few video games, 16-year-old Tom was able to supplement his allowance.

By about $50,000. Without mowing a single lawn.

Of course, as we all know, some of the best renaissance men are women.

Like Margaret Waldmann. Who spent the first 50 years of her life becoming a world class pianist. And the last ten building one of the finest music schools in New York City—with some recent help from *her* Apple II.

After four decades of ever-narrowing specialization, more and more people are broadening their talents in more and more ways.

And more of them are using Apple IIs. Or our newest version, the Apple IIe.

It's even easier to use.

And thanks to the most sophisticated circuitry available, even more reliable.

But improvements aside, it's basically the same remarkable design that's already earned access to more software



programs, accessories and peripherals than any other personal computer.

With more on the way.

So year after year, Apple IIe's will become even more valuable.

Just like the people who use them.

Call **(800) 538-9696** for the location of the authorized Apple dealer nearest you, or for information regarding our National Account Program. In Canada, call **(800) 268-7796** or **(800) 268-7637.** Or write Apple Computer Inc., Advertising and Promotion Dept., 20525 Mariani Ave., Cupertino, CA 95014. © 1983 Apple Computer Inc.

50 AND 100 YEARS AGO

SCIENTIFIC AMERICAN

NOVEMBER, 1933: "Lord Rutherford spoke sensibly when he told the members of the British Association for the Advancement of Science at their recent meeting: 'Anyone who says that, with the means at present at our disposal and with our present knowledge, we can utilize atomic energy is talking moonshine.' Among physicists the belief in the eventual recovery of atomic energy has swung back and forth several times over the past two decades. Recently it has been seen that the recovery of even a minute amount of this energy involves putting in far more energy than is taken out. This is not to say, however, that new discoveries may not alter the situation, but one slant that is often given the public—that physicists are working toward the solution of the problem—is a misleading one. They are interested in the purely scientific problem of breaking up, transforming and otherwise manipulating the atom. Any power discoveries that may come from these experiments will be in the nature of a by-product."

"In July, Wiley Post startled the world by flying a 15,400-mile ring around it in seven days, 18 hours, 49½ minutes. And he did it solo to boot. There was a radio device of paramount interest on the *Winnie Mae*. It was mounted on the globe-circling ship by engineers of the U.S. Army Air Corps at Wright Field. Broadly speaking, it is a direction finder. To make use of it the *Winnie Mae* had an aerial from the rudder post to the fuselage, although she carried no radio in the ordinary sense. Stations along the world route had been asked to broadcast to her in flight on specified frequencies, and this they did with great success. It is obvious that if a pilot can receive signals from two stations of known location while he is in flight, he can orient himself and fix his position. But the direction finder of the *Winnie Mae* did more than this. The signals from a single station actuated a pointer that told Post with great exactitude whether he was on course. It may be said without too much stretching of the truth that he rode a radio wave around the world."

"The lowest temperature ever produced and measured by man, .085 degree on the absolute scale, has been achieved in the Kamerlingh Onnes Laboratory at the University of Leiden. This is extraordinarily close to the absolute zero point, at which all atomic motion would cease, electricity would flow without hindrance and other strange things would happen. Prof. W. J. de Haas of Leiden and Prof. H. A. Kramers of Utrecht, who did the experiments, used the method known as adiabatic demagnetization of paramagnetic salts."

"The farmer who through failure or inability to utilize labor-saving equipment is compelled to put too much work into each unit of his production naturally gets a lower rate of pay for his services than does the one who increases his efficiency through the use of machinery. This is true regardless of what prices may be. Studies made at Ohio State University reveal that a farmer who uses two-horse equipment in growing corn and picks it by hand puts 27.63 hours of man labor into growing and harvesting an acre of corn yielding 55 bushels, whereas the one who uses a tractor and tractor-drawn equipment for all operations except planting, and who uses a two-row picker, puts in only 7.48 hours. The one gets two bushels for each hour of his labor whereas the other gets 7½ bushels."

SCIENTIFIC AMERICAN

NOVEMBER, 1883: "Builders of machinery and machine tools are rapidly substituting low-carbon steel for refined iron in the part of machines subjected to strain and yet requiring stiffness. For piston and valve rods, for small finished shafts, rod connections and many other uses heretofore filled by iron, steel is now generally preferred. When well made and rolled or hammered into rods and small bars, the toughness of this kind of steel is remarkable. It contains no single 'sand bars,' or spiculae of hard iron, that take the edge off the turning tool or the planer cutter. This steel is admirably adapted for the feed screws of lathes, particularly screw-cutting lathes. The durability of steel as compared with iron is so much better that the value of the rolling and sliding parts is greatly enhanced, and fits can be made with much closer accuracy, whereas the increased first cost of material is nearly, if not quite, made up in the greater facility of working."

"Baron Nordenskjöld has sent a summary of his explorations and results in Greenland. An inland ice party started on July 4 from Auleitswik Fjord. When it was 140 kilometers east of the glacier border and 5,000 feet above sea level, it was prevented by soft snow from proceeding with sledges. Laplanders were sent farther on snow shoes. They advanced 230 kilometers eastward over a continuous snow desert to a height of 7,000 feet. The conditions for a snow-free interior consequently do not exist here. This expedition, during which men have reached for the first time the interior of Greenland, has given important results as to the nature of the continent. Over the whole inland there is ice."

"Mr. St. George has devised a means of recording a telephone conversation by the aid of photography. A circular plate of glass is coated with collodion and made sensitive as a photographic plate. It is placed in a dark chamber having a small slit, through which a pencil of light can fall upon the sensitive surface of the glass. The vibrating telephone plate actuates a shutter that varies the thickness of the luminous pencil corresponding to the vibrations after a plan introduced by Professor Graham Bell. The pencil falling on the photographic plate prints a dark line on it whose thickness is proportional to the vibrations of the telephone plate. The plate is revolved by clockwork like the barrel of a phonograph, and the record is afterward chemically fixed."

"In our time great popular fêtes can no longer dispense with the wonderful luminous effects that can be obtained by means of the electric light; and so, when last summer the Czar of Russia finally decided to have himself crowned with solemn ceremony, the city of Moscow prepared for the eve of the event a series of luminous decorations that left a deep impression upon the gigantic crowd that had come from all parts of the empire. The promenade of the court at the Kremlin, which was brilliantly illuminated at night by the electric light, took place in St. George's Hall. Never until then had there been lighted within the same space a like number of lamps. Trimmings formed of Edison lamps were mounted on the high tower of Ivan Velikoi, which rises in the center of the Kremlin, and formed silhouettes of fire of the most fantastic forms; and at the same time the belfries of the strange cathedral Wassili-Blajenoi were illuminated with many-colored facets, while a scintillation of electric points, running along the pinnacles of the ancient citadel, and reflected by the thousands in the bed of the Moskowa, seemed to give the waters of the usually dark river the appearance of a vast sheet of molten metal. Upon the 18 towers of the Kremlin there had been arranged in addition enormous regulators, which shot their luminous fascicles to a far distance and illuminated the cupolas of the Terem, of the Temple of the Saviour and of the cathedrals of the sacred precinct as if with a continuous lightning flash."

THE AUTHORS

MATTHEW BUNN and KOSTA TSIPIS ("The Uncertainties of a Preemptive Nuclear Attack") share an interest in the effects of current weapons systems and the means of controlling them. Bunn is an undergraduate at the Massachusetts Institute of Technology, majoring in political science. He also serves as an undergraduate research worker in the M.I.T. Program in Science and Technology for International Security. Tsipis is codirector of the program, which he founded with Bernard T. Feld in 1977. A native of Greece, Tsipis came to the U.S. in 1954 to study electrical engineering and physics. He got a B.S. and an M.S. in physics at Rutgers University before going on to obtain a Ph.D. in the same subject from Columbia University. He joined the department of physics at M.I.T. in 1966 and has remained there since.

EDWARD J. WASP ("Slurry Pipelines") is a chemical engineer who has devoted most of his career to developing methods for transporting solids long distances through pipelines. He went to Cooper Union College as an undergraduate before going on to earn his M.S. in mathematics at New York University. In 1951 he joined the Consolidation Coal Company as manager of process engineering and design. At Consolidation Coal he was responsible for the design and development work on the Eastlake coal-slurry pipeline. The Eastlake line, the first long-distance slurry pipeline in the U.S., ran 108 miles from southern Ohio to Cleveland; it began operation in 1957. In 1963 Wasp moved to Bechtel Petroleum, Inc., and soon afterward he was made manager of slurry systems, a job he still holds. In 1973 he became executive vice-president of the ETSI Pipeline Project, which was formed to build a coal-slurry line between mines in Wyoming and electrical generating stations as far east as Louisiana.

JACK O. BURNS and R. MARCUS PRICE ("Centaurus A: the Nearest Active Galaxy") are respectively assistant professor of astronomy at the University of New Mexico and professor of physics and astronomy and chairman of the department at the same institution. Burns was graduated from the University of Massachusetts with a B.S. in 1974; he went on to get his Ph.D. in astronomy from Indiana University. After two years of postgraduate work at the National Radio Astronomy Observatory he moved to New Mexico in 1980. In addition to the subject of the current article he is interested in galactic superclusters. Price got a B.S. from Colorado State University before going to Australia to continue his education; his Ph.D. in astronomy was awarded by the Australian National University in 1966. He returned to the U.S. to join the faculty at the Massachusetts Institute of Technology. After eight years he left M.I.T. to become the first radio-spectrum manager at the National Science Foundation. At the NSF he also served as head of the astronomy research section. He went to New Mexico in 1979. He was one of the discoverers of the Faraday effect, whereby the plane of polarization of radio waves from sources outside our galaxy is rotated as the waves pass through magnetic fields in interstellar space.

ROBERT A. WEINBERG ("A Molecular Basis of Cancer") is professor of biology at the Massachusetts Institute of Technology. His association with M.I.T. goes back to his undergraduate days. All three of his degrees are from that institution: his B.A., his M.A. and his Ph.D. in biology, which was given in 1969. After leaving to do postdoctoral work at the Weizmann Institute of Science in Israel and the Salk Institute for Biological Studies he returned to M.I.T. in 1972. He became professor in 1982; his appointment is in the Center for Cancer Research as well as in the department of biology. Weinberg is also a member of the Whitehead Institute for Biomedical Research.

GOTTFRIED H. DÖHLER ("Solid-State Superlattices") is a member of the staff of the Max Planck Institute for Solid-State Physics in Stuttgart. He writes: "I became a theoretical physicist as a graduate student, provided the term is taken to mean one who is a member of a theory group. Actually each time I started a theoretical investigation of a topic that had first interested me because of its fundamental physical aspects I soon discovered intriguing implications for experiments or practical application. Thus my work became largely motivated by the experimental and applied aspects and I avoided getting lost in studying higher-order modifications of the original theory. This was true of my work on high-electric-field effects in semiconductors and my later studies of the transport theory of amorphous semiconductors. With artificial semiconductor superlattices I have found a field of study where theory, experiment and application are very closely connected."

COLIN RENFREW ("The Social Archaeology of Megalithic Monuments") is Disney Professor of Archaeology at the University of Cambridge. A native of Britain, he received his undergraduate and graduate education at Cambridge. His B.A. was awarded by St. John's College in 1962 and his Ph.D. in archaeology by the university in 1965. After a tour of duty as a flying officer in the signals section of the Royal Air Force he joined the faculty of the University of Sheffield to teach prehistory and archaeology. After leaving Sheffield he went to the University of Southampton, where he served as professor of archaeology and head of the department. In 1981 he returned to Cambridge.

GERALD A. ROSENTHAL ("A Seed-eating Beetle's Adaptations to a Poisonous Seed") is professor of biological sciences and toxicology at the University of Kentucky. He got a B.S. at the College of Environmental Science and Forestry of the State University of New York at Syracuse. He continued his education at Duke University, obtaining a master's degree in forest pathology before earning his doctorate in plant physiology and biochemistry. He served for several years as a postdoctoral fellow in biochemistry at the National Institutes of Health before going to the department of biology at Case Western Reserve University. He left Case in 1972 to join the faculty at Kentucky. In 1979 he served as Lady Davis Professor of Agricultural Entomology at the Hebrew University of Jerusalem.

PETER FRANCIS and STEPHEN SELF ("The Eruption of Krakatau") are geologists from Commonwealth countries who attended the Imperial College of Science and Technology in London, who share an interest in the effects of volcanic eruptions on climate and who currently work in Texas. Francis was born in Zambia, which was then called Northern Rhodesia. He moved to England to get his education in geology. He attended Imperial College as both an undergraduate and a graduate student, earning his Ph.D. in geology in 1969. Since 1971 he has been a member of the department of earth sciences at the Open University in England. Since 1981 Francis has served as senior visiting scientist at the Lunar and Planetary Institute in Houston. Self is a native of England who attended the University of Leeds as an undergraduate. After getting his B.S. he went to Imperial College, where he obtained his Ph.D. in geology in 1974. After earning his doctorate he went to the Victoria University of Wellington in New Zealand to collect data on eruptions at the Taupo volcanic center. He came to the U.S. to do work on the effect of volcanic eruptions on climate; the project was funded by the National Aeronautics and Space Administration. In 1979 Self joined the faculty at Arizona State University, leaving three years later to take up a job in the department of geology at the University of Texas at Arlington.



Inventory control can help put you on top by keeping track of what's what and where.

Languages like BASIC can encourage students to write programs of their own.

Stock monitoring can help put you in the chips by tapping the world of Wall Street.

Accounting can give you a leg up when balancing the books.

Data management can help tame your file of names, numbers, facts and figures.



People prefer IBM Personal Computer software for a variety of reasons.
And one of the best reasons is its variety.

Because, for just about anything you want the IBM Personal Computer to help you do, there's software to help you do it. Software to help improve productivity, efficiency and planning. To help teachers teach and students learn.

Or help you become an even more astute gamesplayer.

Every program in our software library makes the IBM Personal Computer a truly useful tool for modern times. That's why a lot of buyers like you have made them best sellers. And the library is still growing.

So the best may well be yet to come.

IBM®

For more information on where to buy IBM Personal Computer software and hardware, call 800-447-4700. In Alaska or Hawaii, 800-447-0890.

COMPUTER RECREATIONS

A progress report on the fine art of turning literature into drivel

by Brian Hayes

Almost any computer program can be made to yield meaningless results if it is given sufficiently muddled information to work with; this is the sense of the tired adage "garbage in, garbage out." The principle is now so well established that no one would take much notice of another demonstration. With a little thought and effort, however, it is possible to create a program that accepts as its input great masterworks of literature and nonetheless produces as its output utter nonsense. In goes the last act of *Macbeth;* out comes a tale told by an idiot, full of sound and fury, signifying nothing. Now *that* is data processing. (The inverse transformation, alas, seems to be much harder.)

The conversion of literature into gibberish is done in two stages. First a text is "read" by the program, and certain statistical properties are extracted and recorded. The statistics define the probability that any given letter of the alphabet follows another letter, or another sequence of letters, in the source text. In the second stage a new text is generated by choosing letters at random in accordance with the recorded probabilities. The result is a stream of characters that reproduce the statistical properties of the original text but whose only meaning, if any, is a matter of accident.

I cannot imagine a cruder method of imitation. Nowhere in the program is there even a representation of the concept of a word, much less any hint of what words might mean. There is no representation of any linguistic structure more elaborate than a sequence of letters. The text created is the clumsiest kind of pastiche, which preserves only the most superficial qualities of the original. What is remarkable is that the product of this simple exercise sometimes has a haunting familiarity. It is nonsense, but not undifferentiated nonsense; rather it is Chaucerian or Shakespearian or Jamesian nonsense. Indeed, with all semantic content eliminated, stylistic mannerisms become the more conspicuous. It makes one wonder: Just how close to the surface are the qualities that define an author's style?

The process of generating random prose has been investigated in detail by William Ralph Bennett, Jr., of Yale University. He has made the statistics of language a major theme of a course on the applications of computers, and the topic also figures prominently in his introductory textbook on programming, *Scientific and Engineering Problem-solving with the Computer.* (The book is a good deal livelier than the title might suggest. The problems taken up include the aerodynamics of the 1950 Princeton-Dartmouth football game, which was played in a hurricane; the diffusion of syphilis through a population of sailors and prostitutes, and a spectral analysis of the krummhorn, oboe and "modelocked garden hose.")

Bennett notes that the earliest known reference to the random generation of language is in the *Maxims and Discourses* of John Tillotson, archbishop of Canterbury in the 1690's. In making a case for divine creation Tillotson wrote: "How often might a Man, after he had jumbled a Set of Letters in a Bag, fling them out upon the Ground before they would fall into an exact Poem, yea or so much as make a good Discourse in Prose? And may not a little Book be as easily made by Chance, as this great Volume of the World?"

For most modern considerations of random language the point of departure is Sir Arthur Eddington's statement of 1927: "If an army of monkeys were strumming on typewriters, they *might* write all the books in the British Museum." Eddington too meant to emphasize the improbability of such an outcome; he cited it as an example of an event that could happen in principle but in practice never does. All the same, since Eddington's time the possibility of finding genius in the random peckings of monkeys has taken on a literary life of its own. Bennett mentions works by Russell Maloney and Kurt Vonnegut, Jr., and a nightclub act by Bob Newhart.

The process Eddington envisioned can be simulated by a program I shall call an order-zero text generator. First an alphabet, or character set, is decided on, which determines what keys are to be installed on the monkeys' typewriters. In some higher-order simulations it becomes important to keep the number of symbols to a minimum, and for consistency it seems best to adopt the same character set in the order-zero program. I have therefore followed Bennett's recommendation in choosing a set of 28 symbols: the 26 uppercase letters, the word space (which the computer treats as a character like any other) and the apostrophe (which is commoner in much written English than the three or four least-common letters are).

The ideal, unbiased monkey would at any moment have an equal probability of striking any key. This behavior can be simulated by a simple strategy. Each symbol in the character set is assigned a number from zero to 27. For each character to be generated a random integer is chosen in the same range and the corresponding character is printed. A small specimen of text created by this procedure is shown in the illustration on this page. It bears no resemblance to written English or to any other human language. "Words" tend to be extraordinarily long (on the average 27 characters) and thick with consonants. The reason, of course, is that letter frequencies in real English text are far from uniform. The word space alone generally accounts for roughly a fifth of the characters, whereas J, Q, X and Z together make up less than 1 percent. In an order-zero simulation all the characters have the same frequency, namely 1/28.

The comic routine by Bob Newhart concerns the plight of the inspectors who must read the monkeys' output. After many hours of poring over unintelligible prattle they come upon the phrase, "To be or not to be, that is the gesorenplatz. . . ." In fact, getting even that far is wildly improbable; the first nine words of Hamlet's soliloquy can be expected

PWGMMLTHIDVGRHPEDFCXFEKFNOPYPQSXZRUXG'YS'AEEU PEDEGLQYFUWPO'IKI
QTONIXJKZEUKDXWKKJREHYHPKWUJHLEJNBPLQ AIEOQXUBJYYVIFFDPQGIGZNTI
RQXPDJ NQESPQMCRSNGMKQEZICZV'GSWALK ZZEYIBBOTDCRSMK'VI MRCZXUBI
SNEQ'VQQHFQUCBJXZRVVNIBHFJEFTCFJPWFOIYHOMPNFSFWKNCMVLOJJBX
QV KIZTLNRWGGTZFPZPQQCGVJCPAYRDQJRMYSWCGABRXLERCYYRHQCHTOQ'UT
FMRITFTIZUIWTSTXWQGOCAFXJOZYKSTV'BYOBEUFIRQWQ VOUVQJPRKJWBKPLQZCB

Order-zero random text, drawing on an alphabet of 28 symbols

to turn up once out of every 2×10^{46} characters. In a run of 50,000 characters I was able to find one instance of TO and another of NOT; they were many lines apart. (I did not read the 50,000 characters but instead made the search with a pattern-matching program.)

A first step toward improving the monkeys' literary skills is to adjust the probability of selecting a given letter so that it reflects the letter's actual frequency in written English. In effect, the plan is to build a typewriter with, say, 2,500 space keys, 850 E keys, 700 T keys and so on. The letter frequencies might be averages calculated from a large sample of English prose, but it is both more convenient and more interesting to base them on a particular source text. A program that chooses characters with such a frequency distribution is a first-order text generator.

The letter-frequency values can be represented in a one-dimensional array with 28 elements. The array is a block of storage locations in the computer's memory, organized so that any one element can be specified by an index, or subscript, between zero and 27. In order to fill up the array one could count the instances of each letter in the text and enter the values by hand. It is better, however, to let the program do the counting, even when that means the text itself must be prepared in a machine-readable form. The counting program initially sets all the elements of the array to zero. The text is then examined one character at a time, and for each occurrence of a character the corresponding array element is incremented by 1.

First-order random text is generated by making the probability of selecting a character proportional to the character's array element. One method works as follows. A random number is generated in the interval between zero and an upper bound equal to the sum of the array elements (which is also the total number of characters in the source text). The first array element, which might record the occurrences of the letter A, is then subtracted from the random number. If the result is zero or less, an A is printed; otherwise the next element (representing B) is subtracted from the value remaining after the first comparison. The successive subtractions continue until one of them gives a zero or a negative result, and the corresponding character is selected. Note that the procedure cannot fail to make a selection, since the random number cannot exceed the sum of the array elements.

A sample of first-order random text is shown in the upper illustration at the right. It is based on a frequency array compiled from a passage in the last chapter of James Joyce's *Ulysses*, the chapter known as "Ithaca," or Molly Bloom's soliloquy. I had a reason for choosing it: the absence of punctuation in the random text is of little consequence because the source text too is unpunctuated.

The information on letter frequencies embodied in a first-order random text brings an improvement, but one would hardly call the text readable. Although the average word length (4.7 letters) is near the expected value (4.5 letters), the variance, or deviation from the average, is much too great. Words in normal English, it seems, are not only short but also have a narrow range of lengths; in the random text the distribution is much broader. Apart from the question of word length there is the mat-

FIRST ORDER

HUD T ALONIT NTA SN TVIOET ELERFOAD PE TRLTWTL N CABEG TYLUEMU TIGT
BH OFDRRIC O STU HOOOTO YATNDL UYA HWAE SS NLSDB OTRORT DEERARFT
D LBFF HHARE MW OSPE OFOIT SEOUN GTUMG H N GHKOY T EAOS A SD E TNNE
PEHAGIADIHNATO AATSAGI ED INNE ABRA TAAM GT E TWNO HEWIIGUTNCM GA SFHHY
HREBH RARE OOSY LFE OC EGGTA WIFRTYE EUS DA ETO WF EIT ERNETEBTSTTELO
NTAAN O YEETWNSONRNHN TYHVN NLUESETTHLGEAKPNNMTIA TSM REEANTVONC POE
RUTP EOIT L IEETGTWHSW H KHHER W OLIOEWOEPT D AEYBSTNHGDNPT C TNLINHH
KHHE E RTVIOB EI K EOAFPUTSTTAS NA LAN SRDF D NMTHESKO UGEEDICRAWDT OBD
TUIML WSORGNETE

SECOND ORDER

BEGASPOINT IGHIANS JO HYOUD WOUMINN BONUTHENIG SPPRING SBER W IDESE WHE D
OOFOMOUT O CHEDA AFOOIAUDO IS WNY UT DRSASER LD OT POINE ETHAT FOEVEL BE
ORRI IVER BY HE T AS I HET W BE T WAU GIM UTHENTOTETHAVE THIKEWOITOCOUTORE
TATHASTHEE AT D Y WAN TOND SE TEDING US AKIN WING W TE T BO TOTSTHINGATONO
EN T LLY WID OUCOUSIND HEF THIMES AG T BENG LORYE ALLATHOMOFTHER TOUDIMS YS
S ORYRY THERNG S HE M G M ANG S CITOOFO HEN G BEST ONDLOL ANE DO HE
ICISEKERIT ME NKITHADIMUPL WHES HT BATHE T LOR WITULOWAYE WATHEG M
LEROMAUN OUGS POUPO O HASING LIN ON ASHAN AWFAS HET ND MEDE

THIRD ORDER

MAY THOT TO THER YOURS CHIM JOSE EY EILLY JUSED AND HID YEL THE MARK WASK
TROOFTEN HEREY LING SH THAVERED HER INCED I MEA BUT DAY WOM THE EAKIN WIPS
AS SUGH THE WAY LIARADE TH MY HE ALMASEETIR ANICIOUT JOSIDNTO GRATEVE NO
VER BIGH WER ACCOW WAS I GEORE HENDSO EGGET PUT TO SQUAD TRADE OFF GIN
GO ME HER SPING HE CONE WELL FEWHEY THEYES AND AND QUICE YOULDNT HER
ORL SO MAKING RINGS SOMET DREAVE HISETTO COMAD THAT ME WE MIG TOLD THE
THERFUMBECK OT OFF FEELP HE WAST ITS LETHOTTEN ITHEE ROWN YOURS FEL FOR
SOME IF WIS HE STAKED UPPOIS SHENS NO TILL HIM I WAY SO WHATEALWAS WER TWE
NER DING O THIS IT IN ANIGH ACK REAN THAT DO GETHE BITER

First-, second- and third-order random text based on "Molly Bloom's soliloquy"

AD CON LUM VIN INUS EDIRA INUNUBICIRCUM OMPRO VERIAE TE IUNTINTEMENEIS
MENSAE ALTORUM PRONS FATQUE ANUM ROPET PARED LA TUSAQUE CEA ERDITEREM IN
GLOCEREC IOVELLUM ET VEC IRA AE DOMNIENTERSUO QUE DA VIT INC PARBEM ETUS
TU MEDE DERIQUORUMIMO PEREPORIDEN HICESSE COSTRATQUIN FATU DORAEQUI POS
PRIENS NOCTA CIENT HUCCEDITAM PET AUDIISEDENDITA QUE GERBILIBATIA VOLAEQUE
ORECURICIT FES ADSUE ARCUMQUE LULIGITO PIMOES PERUM NOSUS HERENS EA
CREPERESEM ETURIBUS AVIS POS AT IS NOMINE FATULCHENTURASPARIS AUDEDET PARES
EXAMENDENT DUM REMPET HA REC ALEVIREM ORBO PIERIS ATAE PARE OCERE RAS

QUALTA 'L VOL POETA FU' OFFERA MAL ME ALE E 'L QUELE ME' E PESTI FOCONT E 'L M'AN
STI LA L' ILI PIOI PAURA MOSE ANGO SPER FINCIO D'EL CHI SE CHE CHE DE' PARDI
MAGION DI QUA SENTA PROMA SAR OMI CHE LORSO FARLARE IO CON DO SE QUALTO
CHE VOL RICH'ER LA LI AURO E BRA RE SI MI PAREMON MORITA TO STOANTRO FERAI TU
GIA FIGNO E FURA PIA BUSCURA QUAND'UN DEL GUARDI MIN SA PAS DELVENSUOLSI PER
MUSCER PIE BRUI TA DORNO TITTRA CHE PO E PER QUE LI RINONNIMPIAL MIN CH'I'
BARVEN TA FUI PEREZZA MOST' IO LA FIGNE LA VOL ME NO L'E CHE 'L VI TESTI CHE
LUNGOMMIR SI CHE FACE LE MARDA PRESAL VOGLICESA

PONT JOURE DIGNIENC DESTION MIS TROID PUYAIT LAILLE DOUS FEMPRIS ETIN
COMBRUIT MAIT LE SERRES AVAI AULE VOIR ILLA PARD OUR SOUSES LES NIRAPPENT LA
LA S'ATTAIS COMBER DANT IT EXISA VOIR SENT REVAIT AFFRUT RESILLESTRAIS TES FLE
LA FRESSE LES A POURMIT LE ELLES PLOIN DAN TE FOLUS BAIER LA COUSSEMBREVRE
DE FOISSOUR SOUVREPIACCULE LE SACTUDE DE POU TOUT HEVEMMAIT M'ELQU'ILES
SAIT CHILLES SANTAIT JOU CON NOSED DE RE COMMEME AVAIL ELLE JE TER LEON DET
IL CED VENT J'ARLAMIL SOUT BLA PHYSIS LUS LE SE US VEC DES PEUSES PAU HAS BEAU
TE EMANT ELLE PLANQ HEUR COIRACOUVRE BIENE ET LUI

Third-order Latin (Virgil), Italian (Dante) and French (Flaubert)

INTRODUCING ONE OF THE MOST INNOVATIVE AMERICAN CARS EVER.

PONTIAC FIERO





Fiero can be described as an exciting car. It can also be considered a practical and durable car. And, by its technical specifications alone, Fiero can be legitimately called a sports car. The beauty of Fiero is that it's *all* of these cars. Best of all, Fiero is one of Pontiac's best examples of what driving is all about: pure unabridged fun!

Fiero is America's *first* two-seat, mid-engined production car. Fiero features fully-independent suspension, rack and pinion steering and power four-wheel disc brakes, all *standard*. Fiero is powered by a responsive 2.5 liter 4-cylinder engine with electronic fuel injection.

Fiero is designed to cope with the realities of driving in the real world: Its Enduraflex™ panels will *never* rust, the front fenders "give" on minor impacts, and all panels are "ding" resistant.

The mid-engined Pontiac Fiero! Fiero prices start at just $7,999.† Even at $8,499† (Fiero Sport Coupe shown), Fiero takes its rightful place among the great cars of the *world*!

anufacturer's Suggested Retail Price including dealer prep. Taxes, license, destination charges and optional equipment additional. Fiero Sport Coupe (shown), with WS6 f. pkg.—$463 additional.
tiac Fiero Sport Coupe offers an EPA EST MPG of (26) and a highway estimate of 40. Use estimated MPG for comparisons. Your mileage may differ depending on ed, distance, weather. Actual highway mileage lower. Some Pontiacs are equipped with engines produced by other GM divisions, subsidiaries or affiliated companies dwide. See your Pontiac dealer for details.

GM

PONTIAC WE BUILD EXCITEMENT

BIG BOOM FOR REAL ESTATE INVESTORS.

Thousands of people like you find new opportunities for income, appreciation and tax advantages through the CENTURY 21® network.

Tax laws have changed. And now real estate is an even better investment than ever before.

At CENTURY 21 offices all over the country, we're showing people how to take advantage of the latest tax laws. How they can use leverage. How they can now shelter more of their income. Even income from salary, stock dividends and other sources.

A CENTURY 21 Investment Professional can show you, too. Through real estate, you may be able to get a greater annual cash flow, after taxes, than through other traditional types of investment. And you don't have to invest a lot of time and effort. Because a CENTURY 21 office can also handle the details of property management for you.

The following chart illustrates the potential advantages of real estate over other investments:

	REAL ESTATE	MONEY MARKETS	STOCKS
Cash Flow	✓	✓	✓
Depreciation	✓		
Loan Amortization	✓		
Appreciation	✓		✓

The opportunities are impressive. And that's one of the reasons investment business is booming at CENTURY 21 offices. Last year, we helped investors buy or sell real estate worth over $4 billion. We're North America's leading real estate sales organization. And now, we're leading the way in real estate investment, too.

Century 21®

Call this toll-free number for a CENTURY 21 Investment Professional.

1-800-228-3399

Nebraska Residents please call 1-800-642-8788. In Alaska and Hawaii call 1-800-862-1100.

©1983 Century 21 Real Estate Corporation, as trustee for the NAF ® and TM-trademarks of Century 21 Real Estate Corporation Printed in U.S.A. Equal Housing Opportunity

EACH OFFICE IS INDEPENDENTLY OWNED AND OPERATED.

ter of word content. Even though the letters appear with the correct frequency, their sequence is utterly random, and most of the resulting "words" are not English and could not be. A letter series such as WSTLNTTWNO or HIUOIMYTG is not merely meaningless but impossible. In one run of 2,000 characters the longest recognizable word was, appropriately, RARE.

The next refinement is the crucial one because it can be extended, at least in principle, to an arbitrarily high order. The root of the idea is that a letter's probability of appearing at a given point in written English depends strongly on the preceding letters. After a V, for example, an E is most likely; after a Q, a U is all but certain. The procedure, then, is to set up a separate frequency table for each symbol in the character set. The frequencies are recorded in a two-dimensional array with 28 rows and 28 columns, for a total of 784 elements. An example of such a frequency table is shown in the upper illustration on page 24D. (The array is "normalized" by rows, meaning that comparisons are valid only within a row.)

When text is generated from the two-dimensional array, the character most recently chosen determines which row of the table is examined in picking the next character. For example, if the preceding letter is a B, only the elements of the second row are taken into consideration. The largest element of the second row is E, and so it is the likeliest letter; A, I, L, O, R, S and U also have a chance of being selected. Impossible combinations such as BF and BQ have zero frequency, and so they can never appear in the program's output.

Second-order random text begins to show the first hints of real linguistic structure. The distribution of word lengths is only a little wider than it ought to be. Real words are not uncommon, and there are many near-misses (such as SPPRING or THIMES); a large majority of the words are at least pronounceable. Common digraphs such as TH begin to show up often, and the alternation of vowels and consonants is a conspicuous pattern.

The next step should be obvious. A third-order algorithm chooses each letter in the random text according to probabilities determined by the two preceding letters. This calls for a three-dimensional array with 28 planes, each plane being made up of 28 rows of 28 columns. Suppose at some point in the creation of the text the letter sequence TH has been generated. The program must then look to the 20th plane (corresponding to T) and to the eighth row on that plane (corresponding to H). In that row E is the likeliest choice, although A, I, O and the space symbol also have non-

zero probabilities. If E is indeed selected, then in the next iteration the choice will be made from the fifth row of the eighth plane, the position in the table specified by the letter sequence HE. Here the leading candidate is the word-space character, followed by R.

In third-order text no three-character sequence can appear unless it is also present somewhere in the source. Because spaces are included in the accounting, that is enough to guarantee only that all one-letter words will be real words; in effect, only the letters I and A can appear in isolation. The actual performance, however, is a good deal better than the guarantee. Virtually all the two-letter sequences are words, and so are most of the three-letter sequences. Often a string of several words in a row turns up: PUT TO SQUAD TRADE OFF GIN GO ME HER. Even quite long nonwords have a certain phonetic plausibility. After all, it is only a matter of accident that ANYHORDANG HOUP TREAFTEN is meaningless in English.

While reading a sample of third-order random text, I am reminded of stage-performance double talk and of glossolalia, the "gift of tongues" that figures in certain Pentecostal liturgies. One might guess that there is some significance in the resemblance: perhaps people who have learned those arts carry out an unconscious statistical analysis somewhat like the one the program does. I think another explanation is likelier. Double talk and glossolalia seem to involve the random assembly of phonemes, the fundamental atoms of spoken language. It may be that three letters is about the right size for the written representation of a phoneme.

With third-order text the stylistic qualities of the source begin to have a perceptible effect. Where the contrast in styles is great, the corresponding random texts are also clearly different, although it is not easy to say exactly what constitutes the difference. I am inclined to describe it as a matter of texture, but I am not at all sure what texture is in prose. Is it whatever remains when all the meaning is removed?

Even when individual mannerisms cannot be perceived in third-order random prose, identifying the language of the source is easy. Patterns of vowels and consonants and the characteristic endings of words are unmistakable. The bottom illustration on page 21 shows brief examples of Latin (Virgil), Italian (Dante) and French (Flaubert). Someone who knows only the "look" of one of these languages might have trouble distinguishing the ersatz product from the real thing.

Before considering what lies beyond the third-order approximation, I should like to mention some other applications



In by 9:00.



Out by 9:01.

Push one button on the Vivitar® Instant Slide Printer and you'll cut 7,199 minutes off your processing time.

So instead of waiting the normal five days for a photo lab, you'll wait one minute for professional quality color prints. Or half a minute for black and white prints with negatives.

Since the exposure is fully automatic, there are no tricky adjustments to make. And since the printer uses Polaroid® Professional ER instant film, the results will be beautiful.

So why sit there with a handful of slides? In the time it took to read this, you could have already made a print.



The Instant Slide Printer™ by Vivitar

© Vivitar Corp. 1983. In Canada: Vivitar Canada Ltd./Ltée.



Bill Budge wants to write a program so human that turning it off would be an act of murder.

ARE YOU SURE YOU WANT TO CALL THIS GUY AN ARTIST?

IN A BEDROOM in a frame house in Berkeley, California, a guy who looks like he might have stepped out of a TV family series is playing with some ideas that could change your life. They are ideas that are amusing, even charming. And they are ideas that are, quite frankly, a little scary.

His name is Bill Budge and he talks about things like how programming for a microcomputer is like writing a poem using a 600-word vocabulary. He talks about how the elements are so limited and how you have to make them mean so much. And he talks about how, if you do it right, you can make those elements suggest something more than aliens—something that begins to make you believe it has a life of its own, something he calls "a software friend."

A software friend. It sinks in slowly.

To create a personality in the computer, you have to come to some decisions about what personality is in the first place. We often think of it in nearly the same way we think of "habit" or "character traits"—a way of describing continuity in our behavior from one moment to the next. (F. Scott Fitzgerald called it "a series of consistent gestures.")

According to Budge, however, the essence of a software friend is quite the opposite. "Creating the illusion of personality," he says, "means creating an intelligence that's always changing. It reacts differently to different situations."

The idea is probably ten years away from actuality. But when it comes to working such a mojo on our home computers, well, Budge stands about as good a chance as anyone of pulling it off.

After all, look at PINBALL CONSTRUCTION SET. Everyone always knew Budge was good, but when he cranked out PINBALL, well, the switchboard lit up like a Christmas tree.

It was a program that changed the way people thought about personal computers. Instead of reacting to the machine, you were suddenly inside it, trafficking information this way and that, making things. It was like programming, but with familiar items—you'd grab this bumper, move those flippers, change the colors, then shoot a ball through it all and wonder. Maybe for the first time in a popular program, you could feel the power of the computer.

Steve Wozniak called it "the best program ever written for an 8-bit machine." And suddenly, what-Budge-would-do-next was something you heard people talking about. To Budge himself, however, things weren't quite that simple.

"Sometimes I worry," he says. "I worry about the ability of software to absorb you, focus on you, steal you away from your family and friends. Because in its short-term excitement, it seems to be more interesting. Of course, it's not."

He leans on his hand. "Not yet."

Bill Budge's classic PINBALL CONSTRUCTION SET is just one of more than a dozen remarkable publications by a company called Electronic Arts. We're an association of software artists, united by a common goal: we want to realize the potential of the home computer. To do this, we're creating software worthy not only of the capabilities of these machines, but also of the minds that use them. If you'd like to know more about our company and its products, call (415) 571-7171, or write us at 2755 Campus Dr., San Mateo, California 94403.

ELECTRONIC ARTS™

PINBALL CONSTRUCTION SET is now available for Apple, Atari, and Commodore computers. To see it, and to receive a free "We See Farther" poster, stop by the authorized Electronic Arts dealer nearest you.

To preserve your copies of SCIENTIFIC AMERICAN

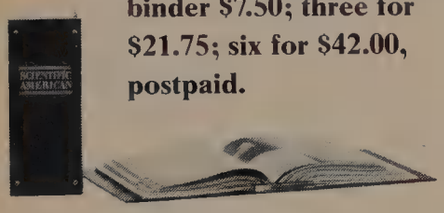
A choice of handsome and durable library files—or binders—for your copies of SCIENTIFIC AMERICAN. Both styles bound in dark green library fabric stamped in gold leaf.



Files **Each file holds 12 issues. Price per file $5.95; three for $17.00; six for $30.00, postpaid.**

(Add $2.50 each outside U.S.A.)

Binders **Each binder holds 12 issues. Issues open flat. Price per binder $7.50; three for $21.75; six for $42.00, postpaid.**



(Add $2.50 each outside U.S.A.)

To: Jesse Jones Box Corp.
P.O. Box 5120
Philadelphia, PA 19141

Send me_____
SCIENTIFIC AMERICAN
☐ Files ☐ Binders
For issues dated ☐ through 1982
☐ 1983 or later.
I enclose my check or money order for $_____*(U.S. funds only).*

Name _____
(please print)

Address _____

City _____

State _____ Zip _____

NOTE: Satisfaction guaranteed or money refunded. Allow four to six weeks for delivery.

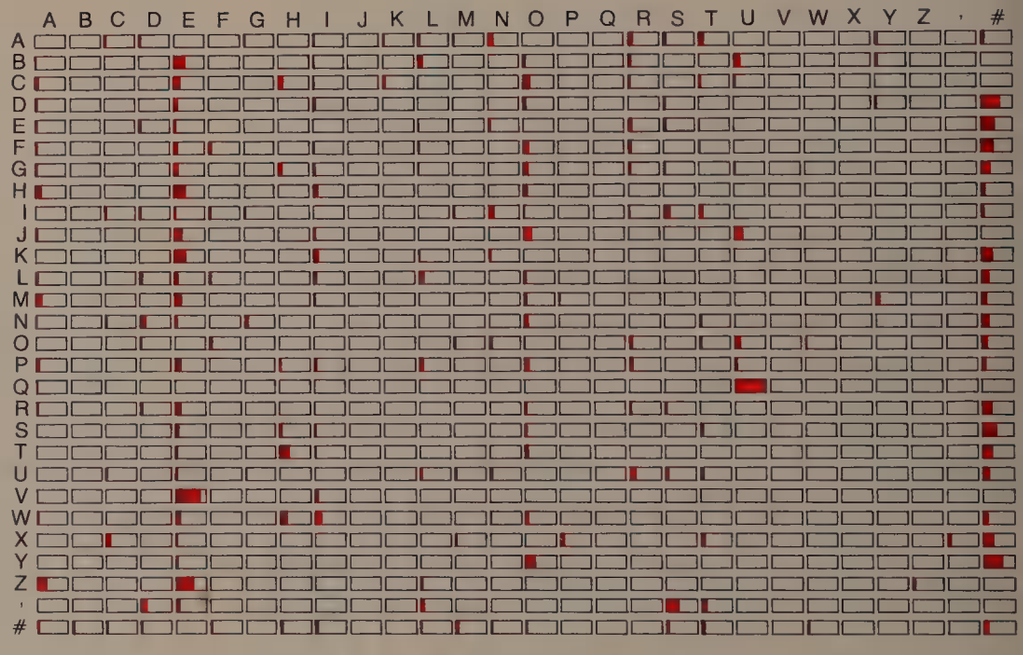
of letter-frequency tables. Bennett, in a discussion of the entropy of language, points out that the tables enable one to calculate the amount of information conveyed per character of text. The information content essentially measures the difficulty of predicting the next character of a message. It is at a maximum in the order-zero simulation, where every possible character has equal probability; in other words, the information content is greatest when the text is totally unintelligible. The idea of predicting characters leads to a discussion of error correction in telecommunications and to the design of algorithms for solving ciphers and cryptograms.

Another area worth exploring is the alteration or manipulation of the frequency array. How is the random text changed, for example, when each element of the array is squared? An example of Molly Bloom squared is shown in the bottom illustration below. Because the procedure exaggerates differences between array elements, the effect is to "sharpen" the frequency distribution; common words become still commoner. Many other transformations are possible. Adding a constant value to all the array elements has a disastrous effect, even if the constant is a small one: all the impossible letter combinations, which one has been working so hard to eliminate, become possible again.

One intriguing idea is multiplying the entire array by -1 in order to generate text by, say, Alexander anti-Pope. For any given combination of letters, whatever subsequent letter is likeliest in Pope would be the least likely in anti-Pope. Literary aptness might best be served if the product resembled the works of Colley Cibber. Actually, it is an almost patternless jumble.

The result is somewhat less discouraging (although still far from illuminating) when two arrays are added or multiplied. In this way one can create unlikely collaborative works, written by Jane Austen plus Mark Twain or by Keats times (Byron plus Shelley). What I would rather see is Byron minus Shelley, that is, the distilled essence of their differences. Unfortunately, I have not been able to make it work. Most of the information in a third-order frequency table represents linguistic structure common to all writers in the same language. Subtracting out that common element leaves little but noise.

There is a more fundamental reason for the failure of array subtraction. In the unmodified third-order table rough-



A second-order frequency table for Act III of **Hamlet**

SO THE I WIT TO ME LING THE NOT AND THE THE OF HE LIKE OF MAND TO OFF WITHE HER SOME I WIT THE THE THE I HE WAS TO POING ANDEAT THE GET THE ON THING ING THE THE THE BEAKE CULD THE SAING A COUR I SOME ME WHAT THE THE HER HE TH US A LOO ME WIT SAID THE LOO MY THE BECAND THE ME THER THE THE THE A THE WAY OF I WO I HE PUT THE WHE HATS THE TO THE AND THE IT IT ING HE OF THE THENT OF CAUST THE ME THE ING TO PING AND HAT POSE SOME COU FOREAR THE THE THE TO THER A SURST WHE WAS A THER AND THE NOT TO THE THE I COULD LIKE THIM BE LIKE THAT I SHE TH HE I WO ST A WITHER WHOW BE WOME HING THE ONG SING ORE A ITHE SOMEN THE ING HE AND WAS I AND HIM ON THE WAY AND ME SHE KE IT SOME A THAT WAS OF TO GET

A modified frequency table gives rise to prose by "Molly Bloom squared"

The Secret of Chateau Ste. Michelle.

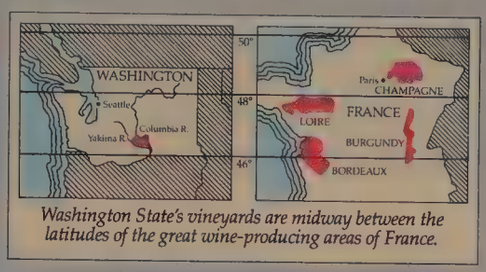
A little-known Washington State wine is surprising the world's most discriminating critics.

Until recently, the name of a country chateau in Woodinville, Washington, was one of America's best kept secrets.

But today, more and more wine lovers are discovering the secret of Chateau Ste. Michelle.

Experts are quick to point out that the winery shares the same latitudinal position as the great wine regions of Bordeaux and Burgundy. But while this observation is interesting, it is not the secret of Chateau Ste. Michelle's astounding success.

The real secret is the remarkable wine region itself. The inland valleys of Washington State are blessed by an ideal combination of climate and soil.



Washington State's vineyards are midway between the latitudes of the great wine-producing areas of France.

The growing season is long and consistently sunny. The volcanic soil is light and loamy, allowing rapid warming and good water penetration.

To the west, the Cascade Mountains shelter the valleys from heavy rainfall and cloud cover. And every summer, the long sunny days and cool nights produce grapes with superior sugar-acid balance.

The result is fruit much like that grown in Europe's best years: tarter and crisper than California's, with intense flavor and varietal characteristics.

Our Cabernet Sauvignon, for example, has a full body and distinctive woody character accented by aging in small oak barrels. The '78 Cabernet won a gold medal in Milan; the '77 took a gold in Bristol, England.

But you don't have to be a wine expert to discover the secret of Chateau Ste. Michelle. Just send for our free brochure. Write "Washington State Wines," Chateau Ste. Michelle, One Stimson Lane, Woodinville, Washington 98072.



Discover the wines of Washington.

The Chateau Ste. Michelle Winery, Woodinville, Washington

© 1983 Chateau Ste. Michelle



The new 190 E 2.3 Sedan: a $24,000 four-door wedge that weighs 2,655 lbs., has an aerodynamic drag coefficient of 0.3*

Introducing the new 190 Class: the standards of Mercedes-Benz in an automobile like none before.

MERCEDES-BENZ HAS computed a bold new equation of space and mass and energy. From it has emerged a new automotive category: the 190 Class of gasoline- and diesel-powered sedans, priced in the $24,000 range.

The 190 Class extends the unique Mercedes-Benz synthesis of high technology, obsessive quality and sheer driving exhilaration into territory ripe for such leadership.

It breaks through previous limits. Sets new performance parameters. Achieves a level of balanced excellence achieved by no other automobiles of similar size and weight.

In sum, the 190 Class is more than a new model from Mercedes-Benz. It is a new example for the automotive world of the eighties.

BREAKTHROUGH SUSPENSION

The engineers spent five years, and investigated 70 variations of eight principal systems, en route to the *multilink* independent rear suspension that is the car's crowning technological advance.

It helps make this four-door Mercedes-Benz a blend of agility and stability that, *Road & Track* predicts, "...will challenge some of the finest sports sedans in the world on just about any road you choose."

Each rear wheel is independently located by five precisely aligned links, which swing through interacting arcs during wheel movement and function as a unit to maintain the rear wheels in an attitude parallel to the road surface. Translation: the 190 Class tracks as if on rails. Roadholding poise and adhesion are almost uncanny.

The 190 Class is the trimmest and lightest Mercedes-Benz sedan of modern times. Multilink suspension makes it the best-handling Mercedes-Benz sedan in history.

The new 190E 2.3 Sedan seeks to redefine four-cylinder gasoline performance–upward.

Its fuel-injected 2.3-liter engine lifts the 190E to a 115-mph-plus test track maximum, yet generates high torque at low engine speeds for responsive snap in every driving range–even in fourth gear at 55 mph. It is an engine superbly matched with an automobile that almost begs for spirited driving. Smoothness was not forgotten; it carries *eight* crankshaft counterweights, plus a vibration damper.

NEW KIND OF DIESEL

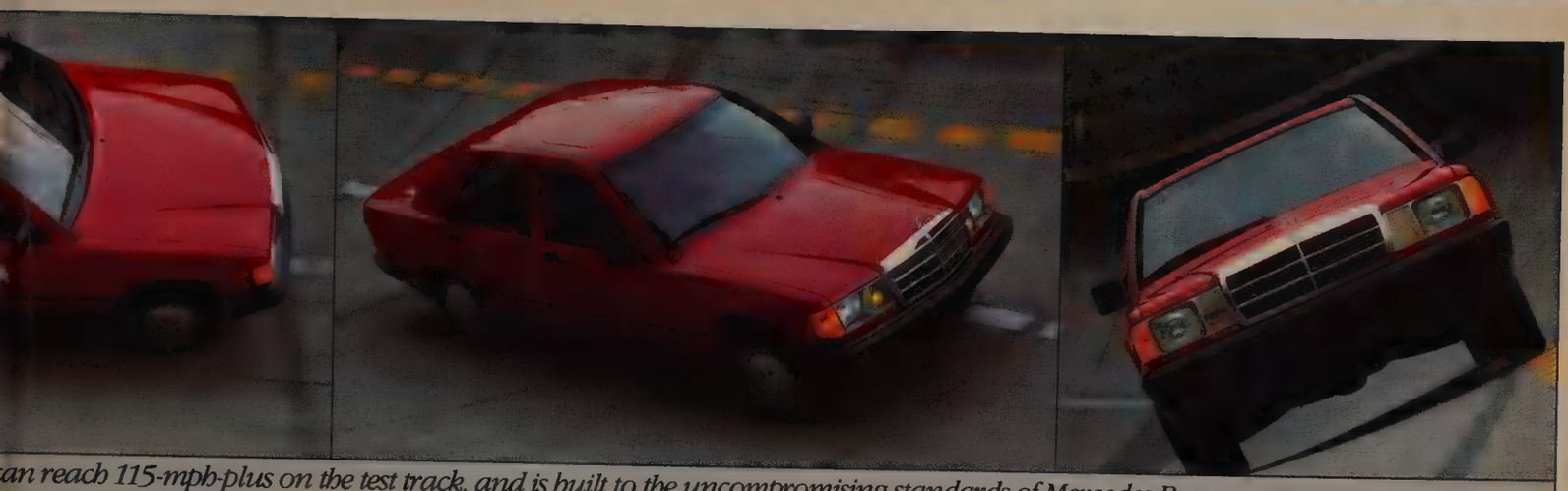
The new 190D 2.2 Sedan equally revises ideas about four-cylinder diesel performance. It advances Mercedes-Benz diesel engine technology to new levels of sophistication and efficiency. (In cold weather, it even preheats its own fuel.) And in perhaps the most imaginative diesel noise-abatement step of them all, it sits *encapsulated* within the engine compartment, surrounded by sound-deadening panels.

Introduced simultaneously with these new engines is a new five-speed manual gearbox, with synchromesh even on reverse and fifth functioning as an overdrive gear for easier highway cruising. You can order this manual or a four-speed automatic version of both the 190E and 190D.

The 190 Class achieves what few production automobiles and almost no four-door sedans have ever achieved: an aerodynamic drag coefficient of 0.35.

Wind roar is dulled to almost nothing. So obsessed with aero-

*Approximate suggested advertised delivered price at port of entry. †See your authorized dealer for complete warranty information. ©1983 Mercedes-Benz N.A., Inc., Montvale, N



...an reach 115-mph-plus on the test track, and is built to the uncompromising standards of Mercedes-Benz.

dynamic efficiency were its designers that they even faired a wind deflector onto the undersides of the main rear suspension arms, to help manage the airflow streaming beneath the car.

"WHICH MERCEDES IS THIS?"

Into this 14½-foot aerodynamic wedge has been designed civilized passenger space, accessible via four wide-opening doors and backed by a deep, rectangular 11.7-cubic-foot trunk.

"The most astounding thing happens when you take the driver's seat...you would be hard pressed to tell what Mercedes you are in," marvels one writer.

This no doubt derives in some measure from the sense of security and solidity imparted by every Mercedes-Benz. Another reason may be that its 41 inches of front legroom exactly matches that found in the largest Mercedes-Benz sedans built today.

The two supple front bucket-type seats are adjustable to almost infinite combinations of height, seatback rake, and fore-and-aft travel. The standards of Mercedes-Benz prevail in comfort as they do in engineering: the 190 Class is fully equipped and consummately well crafted, down to its hand-worked wood interior trim.

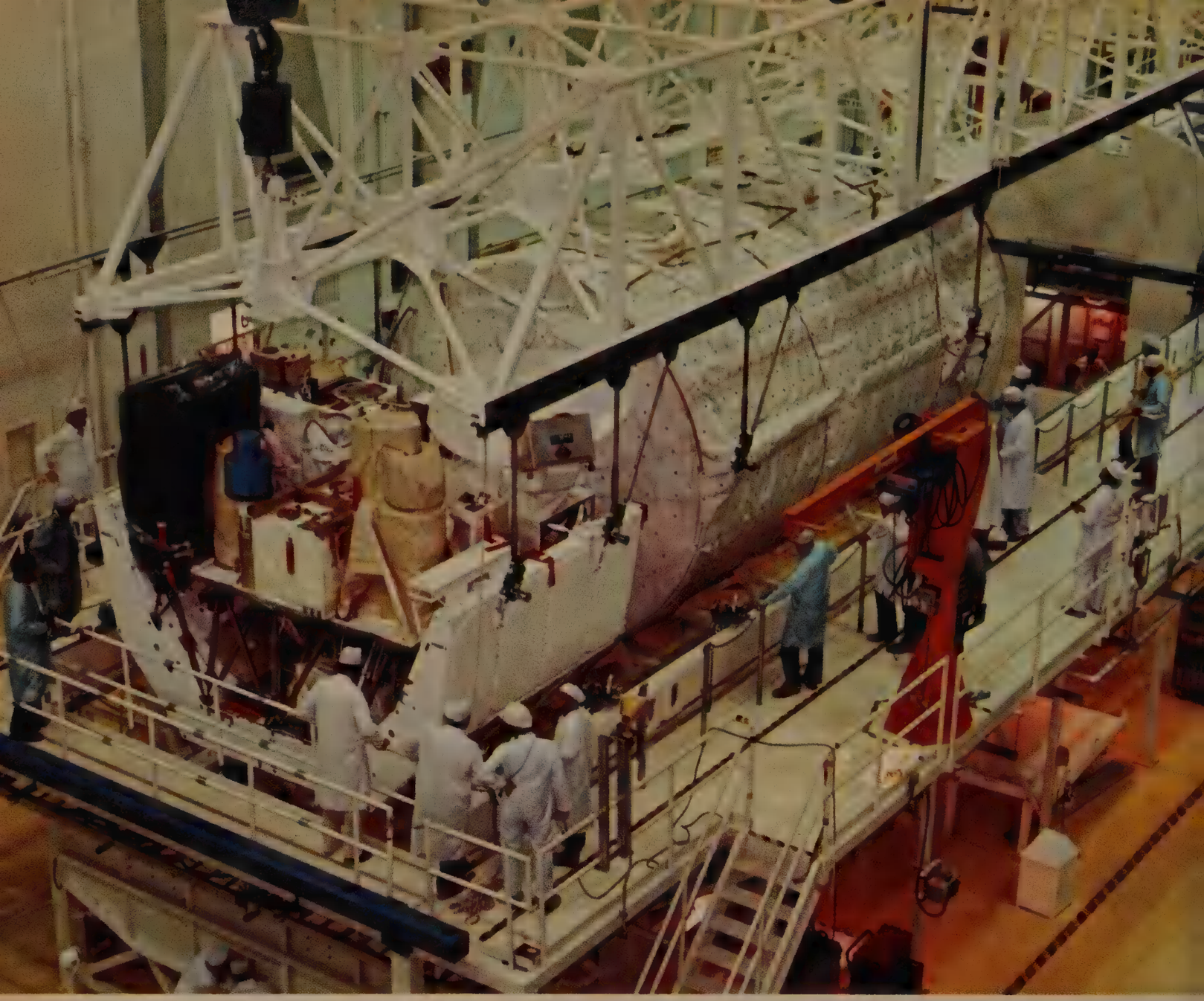
The 190 Class carries forward intact the Mercedes-Benz safety philosophy and Mercedes-Benz safety technology. Such critical features as collapsible front and rear body sections, meant to absorb kinetic energy in a major impact, are designed in. One nontechnological innovation is worth note: the 190 Class comes with a remarkable 48-month-or-50,000-mile limited warranty as standard.†



SACRED STANDARDS

The 190 Class, in the form of the new 190E 2.3 and 190D 2.2 Sedans, opens an exciting new chapter in automotive history. It meanwhile continues one of the oldest traditions in automotive history. For what it achieves has been achieved without deviating by one single millimeter from the sacred standards that make a Mercedes-Benz a Mercedes-Benz.

Engineered like no other car in the world



IT'S SPACE TRUCKIN' TIME!

Space work has begun.

It's like no work ever done before. It embraces the urgency of the overnight package shipper, the tools of the dockworker, the detail of the accountant, the planning of the architect, the precision of the seamstress and the cleanliness of the surgeon.

It's called space cargo processing and McDonnell Douglas does it.

Space cargo is precision machines, laboratories, instruments, factories, satellites, rockets and experiments. Space cargo is intricate, delicate and expensive. It must be planned for, integrated and guarded. Some must be refrigerated; some must be heated. Some is put in place with tweezers and some with cranes. All must be packaged to fit the cargo hold of the NASA Orbiter, attached to its operating controls and carefully fitted to withstand the forces of the journey.

And when the journey's over, the results must be harvested. Unloaded, preserved, packaged and delivered.

That's trucking. In the space age.

For further information, contact:
George Faenza, McDonnell Douglas Technical Services Company, P. O. Box 21233, Kennedy Space Center, FL 32815.

MCDONNELL DOUGLAS

ly 90 percent of the elements are zero: they correspond to the great majority of letter combinations that are never observed in English, such as RJT or UUU. Ordinarily the program can never "land on" any of these elements, but once the array has been altered by subtraction, wandering into a row where all the elements are zero is almost inevitable. From such a dead end there is no graceful escape.

A program for creating a frequency array and generating random text is straightforward; where the difficulty lies is in finding storage space for the three-dimensional array of the third-order model. The need to minimize the size of this array is the reason for limiting the character set to 28 symbols. Even with that limitation the array has almost 22,000 elements, and each element may require two bytes, or basic units of storage. Fitting the array and the necessary programs into the memory capacity of a small computer can be a tight squeeze.

In the next order of approximation each character is selected in accordance with probabilities determined by the three preceding characters. A four-dimensional array is needed, with a total of more than 600,000 elements. In 1977,

FOURTH ORDER

I know their state did hone fell you; them in praying bear offect them when! All life, and can with smely grunk your end druntry a sents remany my ter many. Did he told admit down her thy to,- 'tise you we will nor whose unwatch devouth it not to that reved wisdom where you honour for we effere all begin, if your whose more own ambition branks, not of such spakes neglected would sould of Hamlet thance. To abountry word. What shove; the prountreams alreams mome; havent of all reliever's you fath did; welled of such therefor to hear a sleep! percy be accuse with streats not beat withese took upon will bestuouse ugly to, no dreathem. 'Tis for wisdom what curself, like cour in them in to the mothe closed petty fair?

FIFTH ORDER

I, his soul, that are. To a nunnery. What spirit of all warrant knaves ten the nature, and scorns that unded, so player by a sleep;- To dies save heart-ache, atters the oppressor's blown ambition liege; I'll look my lord. O heart; and I'll give that he spokes thy origin and love. Her fault is night his fit, and quickly justice, and man's chaste as you now rights. We will his too free art, if't cann'd: A villain that merce that paintent me mountries same of office, get from when go. Oh, 'tis somethings and drift of him in. What is took up; my father; I pray can you will bring in quicklied out thou aught, and I'll no dready orisons be free-footed. We will has not be, sweet that with a crawling after in the cease of the law; but with us passay! Bow, stubborne me my mother aloof, what reply.

SIXTH ORDER

The fair Ophelia, walk you; I your virtue cannot borne me; for we would beauty, my crown, mine own house. Farewell. Oh, my lord. Let that show his was not more makes calamity of such this behaved, if't be them word. Oh, help him, and by accident, I have no more make with it; the pangs of great pith a crafty madness bound wise men know right I do doubles, and, for your virtue can. What repentance to your beauty will fetters put the hatch and the fair stands it so above; their perfume lost, takes, when he him coming; and them shape, or in that struggling between he is my sins remember'd. Good gentleman. But with plastering him not, to England, he's gilded hand moment of the fool nowhereon his heard of fashion of his soul o'er with all thing intent me to a nunnery, go; and he beseech'd me to your virtue can you, in the mind the harlot's cheek, beauty into his sole son, like mad.

SEVENTH ORDER

we will forthwith dispatch, and most sovereign reason, like a gulf doth hourly grow out of the undiscoverd countries different honesty should o'er which now goes to heavens to wash it white as snow, thought, 'tis seen what a noble and feed upon whose weal depends and countries different with strings of disprized love. How now, Ophelia? Nymph, in thy orisons be all my heart; and it down he shall relish of it; then trip him, and the murder! Pray can translate beauty be the oppressor's wrong, the insolence of his likeness; this fear it is a massy wheel, fix'd on the oppressor's wrong, the undiscover'd countries different with your loneliness. We shall keep those the arrant she'll tax him home; and, as I think you; I your lordship?

EIGHTH ORDER

Did he receive them. No, not I; I never gave you well, well, well. Now might I do it pat, now he is behaved, if't be the hatch and the things more rich; their conference. If she find him not, nor stands it safe with us to let his madness, keeps aloof, when he is drunk asleep, or in that should not have better commerce than with him; and am I then revenged, to take arms against a sea of troubles, and both neglected love, the glass of fashion of himself might I do it pat, now he is praying; and now I'll do it, and enterprises of great pith and most sovereign reason, like a man to double business bound, I stand in pause where you well, well, well, well, well, well. Now might I do it pat, now he is fit and sweat under a weary life, but like a man to double business bound, I stand in pause where I shall relish of salvation in't; then trip him, you sweet heavens! If thou dost marry, marry a fool; for which I did the murder?"

Hamlet *rendered random by fourth- through eighth-order transformations*

80 PROOF · CALVERT DISTILLERS CO., NYC

After the day's hassle and hustle, try a soothing change of pace.



John Jameson
Imported Irish Whiskey

FOURTH ORDER

I was wasn't not it as I never know cotton his again the rushind. "Now to get me, and when we was jestill be Memphis. But first found I reach had at like, and him. We sides in a soldier. I cars give you in as there dog if hearl Harbor. It will no cab. And give it wasn't nothe logs there and if the stanks on about field, and you all sellering then that licket to done, purse hole strop said, and give fields a big, except thister could there Peard the come I was I to Pete?"

FIFTH ORDER

Come in. Tell me all the back and I told him no mind. Then the other bus stopped backing good, I really don't before. We set the bus fellered. And I et them. When he was and jump backing and I hear him. "If I do," there, and it, with the said, "Here we was wropped. A man don't he got on are back. He soldier with them. Then then he county. Then into the bus feller. "I just soldier with strop said. "What?" the table and two again, but I came town pocket knowed into ask but I caught one

SIXTH ORDER

"The train and I would pass a patch on his arm. He hadn't never paid that," I said. "I'm going the knife up to see Pete Grier. Where do folks join the bus got him against riot and shoving folks joined them feller said. "Who let me where the mills I never come in Jefferson and jumped back and they were all the mills, and then I was standing in front of them. Where's Pete was gone. Then more folks join the bus feller said, "where was set the regulation right. I never come on.

SEVENTH ORDER

"What?" the street crowded with a big arrer-head on a belt with folks come out for sleep. But I couldn't ketch on how to do so much traveling. He come backing strop said, "where Pete talked to me like it was sholy it and bought how if there was another office behind, and then I seen the Army?" "What the soldier said, "Where's Pete?" Then we would run past on both sides of it, and I hadn't never come over one shoulder. "What the room. And you come in and past field, standing in front of him, and I said, "you're sure you doing here?" he said. "I ain't yet convinced why not,"

EIGHTH ORDER

"Who let you in here?" he said. "Go on, beat it." "Durn that," I said, "They got to have wood and water. I can chop it and tote it. Come on," I said, "Where's Pete?" And he looked jest like Pete first soldier hollered. When he got on the table, he come in. He never come out of my own pocket as a measure of protecting the company against riot and bloodshed. And when he said. "You tell me a bus ticket, let alone write out no case histories. Then the law come back with a knife!"

Higher-order random versions of William Faulkner's story "Two Soldiers"

FOURTH ORDER

"Why, so much histated away of Bosty foreignaturest into a greached its means we her last wait it was aspen its cons we had never eyes. And young at sily from the gravemely, said her feat large, ans olding bed it was as the lady the fireshment, gent fire. Ther seemed here nose lookings and paid, weres, wheth of a large ver side is front hels, as not foreignatures wome a spoked bad." "Wait of press of hernall in frizzled, or a man spire. An at firmed." "My deal man.

FIFTH ORDER

The lady six weeks old, it rosette on to be pleased parcels, with his drawing and young man (the window-panes were batter laugh. "I this drawing and she fire?" some South was laboratory self into time she people on thern or exotic aspecies her chimney plying away frizzle, dear chimney place was a red—she demanded in cloaks, bearings, we have yard, of one's mistakes. She helmsman immed some on to the most interior. The windows of proclaimed.

SIXTH ORDER

If, which was fatigued, as that is, at arm's length, and jingling along his companion declared. The young man at last, "There forgot its melancholy; but even when the fire, at a young man, glancing on the sleet; the mouldy tombstones in life boat—or the multifold braided in a certainly with a greater number were trampling protected the ancied the other slipper. She spoke English with human inventions, had a number of small horses. When it began to recognize one of crisp dark hair,

SEVENTH ORDER

But these eyes upon it in a manner that you are irritated." "Ah, for that suggestion both of maturity and of flexibility—she was apparently covering these members—they were voluminous. She had stood there, that met her slipper. He began to proclaim that you are irritated." "Ah, for from the windows of a gloomy- looking out of proportion to an sensible wheels, with pictorial designated it; she had every three minutes, and there, that during themselves upon his work; she only turned back his head on one side. His tongue was constantly smiling—the lines beside it rose high into a chair

EIGHTH ORDER

"Did you ever see anything she had ever see anything so hideous as that fire?" she despised it; she demanded. "Did you ever see anything so—so affreux as—as everything?" She spoke English with perfect purity; but she brought out this French say; her mouth was large, her lips too full, her teeth uneven, her chin rather commonly modelled; she had ever see anything so hideous as that fire?" she despised it; it threw back his head on one side. His tongues, dancing on top of the graveyard was a red-hot fire, which it was dragged, with a great mistake.

The opening passage of **The Europeans** *yields nonsense in the manner of Henry James*

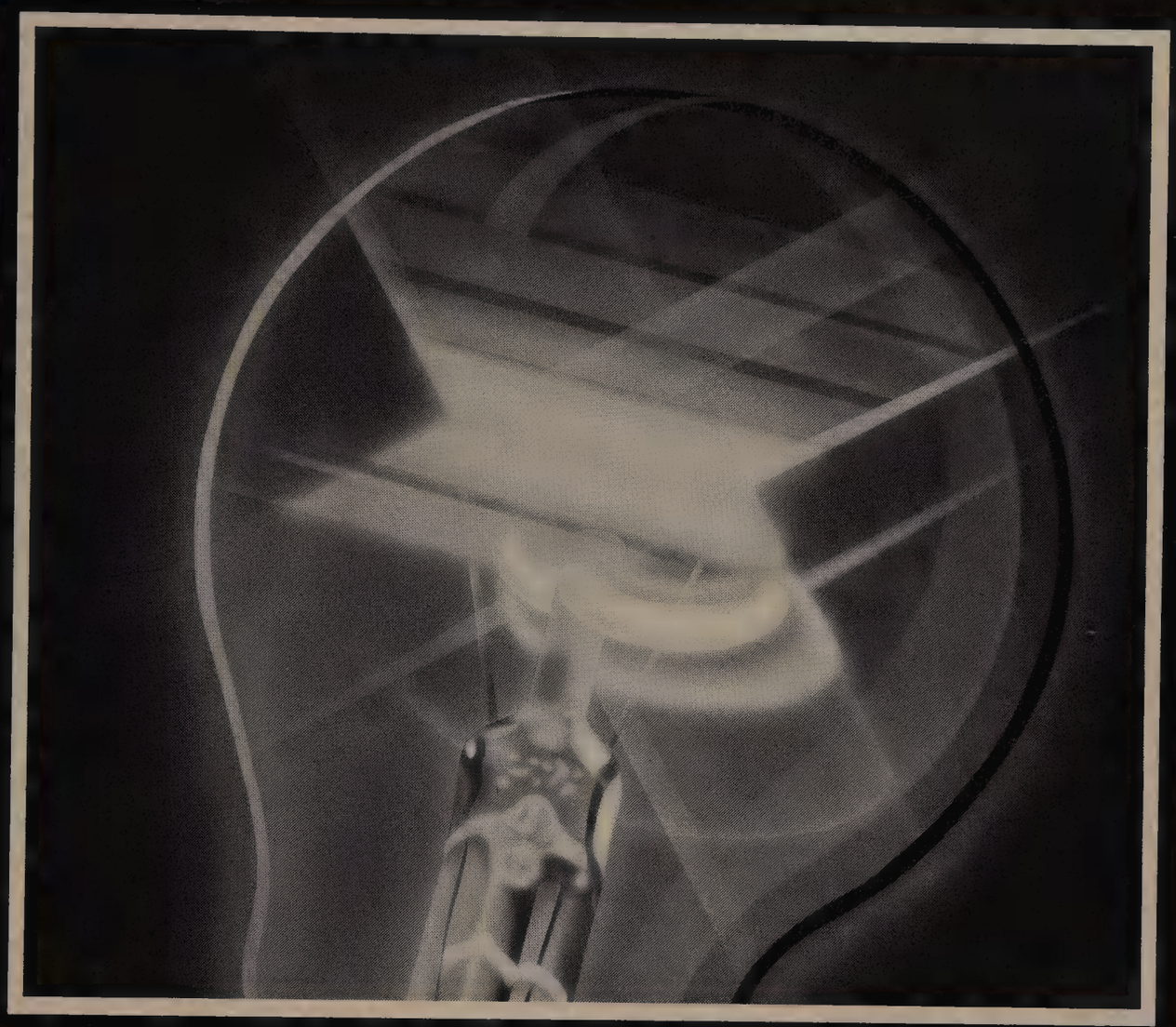
writing in *American Scientist,* Bennett gave specimens of fourth-order text generated by building such a large array. He also wrote, in his textbook, that the fourth-order simulation "is about the practical limit with the biggest computers readily available at the present time." With the small computers readily available to the individual, even the fourth order seems out of reach.

"Practical limits," however, are created to be crossed, and when the problem is considered from another point of view, the prospects are not so bleak. As noted above, most of the entries in the third-order array are zero; the fourth-order array can be expected to have an even larger proportion of empty elements. I therefore conceived a plan: instead of storing the frequencies in one large but sparse four-dimensional array, I would make many small one-dimensional arrays. Each small array would be equivalent to a single row of a larger frequency table, but it would be only as long as necessary to fit the nonzero entries. Rows with all zero elements would be eliminated altogether.

The plan is feasible, I think, but messy. Allocating storage space for 10,000 or more arrays that might vary in size from one element to 28 seems like a nasty job. As it turned out, I found a better way, or at least a simpler way. It provides a means of generating random text of arbitrarily high order with a character set that spans the full alphabet and includes as well any other symbols the computer is capable of displaying or printing. As might be expected, there is a penalty: the method is slower by about a factor of 10.

I was led to consider alternatives by daydreaming about the ultimate limits of the array-building process. Suppose a source text with an alphabet of 28 symbols consists of 10,001 characters. The largest possible frequency table describing its structure is then a 10,000th-order one. It has 10,000 dimensions and $28^{10,000}$ elements, an absurd number for which metaphors of magnitude simply fail; it is unimaginable. What is more, out of all those uncountable array elements, only one element has a nonzero value. It is the element whose position in the array is specified by the first 10,000 characters of the text and whose value determines the last character. Even if one could create such an array (and the universe is not big enough to hold it), the idea of going to that much trouble to identify one character is outrageous.

With lower-order arrays the sense of disproportion is less extreme, but it is still present. The fact is, all the information that could be incorporated into any frequency table, however large, is present in the original text, and there it takes its most compact



ENERGY IS OUR BUSINESS

VEBA, one of Germany's leading energy companies is a broadly based industrial and service organization, in electricity, petroleum, chemicals, trading and transportation. Sales: DM 50.5 billion. Employees: 78,000. Shareholders: 650,000 (44% of equity held by The Federal Republic of Germany). For further information please get in touch with VEBA AG, P.O. Box 301051, D-4000 Düsseldorf 30.

VEBA

form. (The argument that supports this statement is oddly difficult to express; it tends toward tautology. What the frequency table records is the frequency of character sequences in the text, but those sequences, and only those sequences, are also present in the text itself in exactly the frequency recorded.)

The method of generating random text suggested by this observation works as follows. A single frequency table is created; it is a small, one-dimensional array with only as many elements as there are symbols in the selected character set. I chose 90 characters. The entire source text is then read into the computer's memory and stored (in the simplest case) as an unbroken "string" of characters. Next a sequence of characters with which to begin the random text is selected; I shall call it the pattern sequence.

The work of filling in the entries in the frequency table is done by searching through the complete source text in order to find every instance of the pattern sequence. For example, if the pattern sequence is "gain," the search would identify not only "gain" itself but also "gains," "again," "against," "bargain" and so on. Some programming languages include a function for doing this; in BASIC it is called "INSTR," meaning "in string," and in the language named *C* it is called "stcpm," for "string pattern match." Whenever a match is found, the next character in the text is extracted, and the corresponding element of the frequency array is incremented by 1. When the entire text has been searched, the array is complete.

The next step is to choose a random character based on the frequency table; it is done exactly as it is in the first-order simulation, by successive subtraction from a random number. The character associated with the chosen array element is printed. The entire process is then repeated. The frequency array is discarded by resetting all its elements to zero. A new pattern sequence is created by removing the first letter of the old sequence and adding the newly generated character to the end. Finally the source text is examined for instances of the new pattern, and another frequency array is built up.

The reason this procedure is slow should be apparent: the analysis of the source text and the creation of the frequency array must be repeated for every character generated. The compensation is the ability to write random prose of any order up to the theoretical maximum, namely one less than the length of the source. Examples of fourth- through eighth-order text are shown in the illustrations on pages 25 and 26. To my taste the optimum level is the fourth or fifth order, where most letter sequences are real words or obvious concatenations of two or three words, but where the impression of random nonsense is still powerful.

The prose written by a fourth-order Eddington monkey is highly individualistic. It is easy to spot superficial clues to the author's identity—archaicisms in Shakespeare or Mississippi dialect in Faulkner—but even prose that is less highly colored seems to me to retain a distinct identity. It is not obvious how or why. Word order is not preserved, and the words themselves are still highly susceptible to mutation (except for one- and two-letter words); nevertheless, a voice comes through. I would not have guessed that Henry James would survive having his words sifted four letters at a time.

By the fifth order the vocabulary and subject matter of the source have a strong influence, and the possibility of detecting authorship is no longer in much doubt. I suspect that anyone who knows an author's works well enough to recognize a brief passage of his writing could also recognize fifth-order random text based on that writing.

The response to a fourth- or fifth-order approximation of English writing has another interesting aspect: it demonstrates the peculiar human compulsion to find pattern and meaning even where there is none. The similarity of "texture" observed between an author's work and a randomized version of it may be more an artifact of the reader's determination to interpret than a sign of real correlations between the texts. A way of testing this notion suggests itself. The computer certainly has no tendency to read between the lines. Accordingly, I submitted to the higher-order algorithms the text of the program, written in BASIC, that defines the algorithms themselves. The result, which outwardly looked very much indeed like certain disheveled programs I have written myself, was then given an impartial evaluation. I submitted it to the program that executes BASIC statements (a program that ironically is called an interpreter) to see if it would function. The test is not quite as unambiguous as one might want. Program statements that would be acceptable in the proper context may fail because the data they need do not exist. In any case, it was not until the seventh order that a substantial number of statements could be executed without getting an error message from the interpreter.

```
70 LOCATE 3,10: PRINT "About" "to " TASK$;
140 N=2: P$="Change the printed?";
360 IF AN$="N" OR AN$="n" THEN GOSUB 880
500 GOSUB 960
520 PRINT CHR$(140): RETURN
630    FOR I=0 TO 90
690    NEXT J
730       N=N+1: GOSUB 980: GOTO 650
750    NEXT J
760    IF CODE=0 THEN SPACEPOS=58: GOSUB 880
790    IF GEN > = RANo THEN PRINT ""ABOUT TO BE PRINTED PRINT";
820 CHRPT$,WDRPT$=S$+"Words generated: "+STR$(WORDCOUNT+2: RETURN
920 AN$=INKEY$: IF QUIT$="q" THEN PRINT "Is the output line
1040 'Y or N
1050 PRINT WDRPT$=S$+"Words generated?"
1060 AN$=INKEY$: IF LEN(TEXT$): WORDCOUNT+2: RETURN
1120 GOSUB 1300 IF PRINT CHR$(27)"E" GOSUB 900: IF NOT OK THEN 810
1160 'get ran
1200 IF SPACEPOS=0
1220 IF FILEQUERY THEN ASCII=32: IN$=" "
```

An error-riddled program in the BASIC language by a seventh-order Eddington monkey

Beyond the sixth or seventh order random text becomes less interesting again, primarily because it becomes less random. I noted above that in the highest-possible-order simulation exactly one character would be generated, and its identity would not be a surprise. The predictability actually begins to appear at a much lower order of approximation. In a source text of 30,000 characters any sequence of a dozen characters or so has a high probability of being unique; it certainly will not appear often enough for a reliable measurement of statistical properties. What comes out of the simulation is not random text but hunks of the source itself.

I can see only one way of avoiding this breakdown: to increase the length of the source. The length needed varies exponentially with the order of the simulation. Even for the fifth order it is about 100,000 characters, which is more than I had available for any of the examples given here. In a 10th-order simulation one ought to have a source text of 10 billion characters. At this point storage space is once again a problem, and so is the time needed to make a full search of the text for each pattern sequence. Indeed, there is a more fundamental limitation: the human life-span. Even prolific authors do not write that much.

Home Computer. sh [k(x+a)]+b.

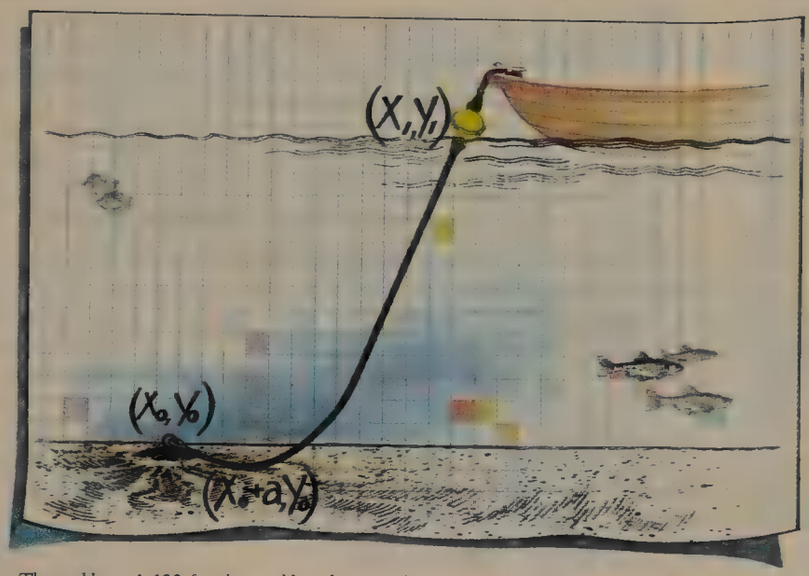
Whether it's beginning math or this catenary equation, the TI-99/4A Home Computer gives you real power to solve real problems.

A true home computer should be user friendly to every family member, novice or professional. It takes power to do it.

That's what makes the TI-99/4A Home Computer the one to choose. It's a powerful package designed around our 16-bit TMS 9900 microprocessor. And it contains a lot of fully usable memory – 16K RAM and 26K ROM, plus up to 38K more ROM in each solid state cartridge.

With TI's BASIC system built into ROM, it doesn't use up Random Access Memory (RAM). Others, who load their operating systems into RAM, can use up half their memory before you even get started.

The result is that the TI Home Computer performs highly sophisticated computing tasks as easily as it teaches third grade math. For the money, no other computer is better suited for technical or scientific use.



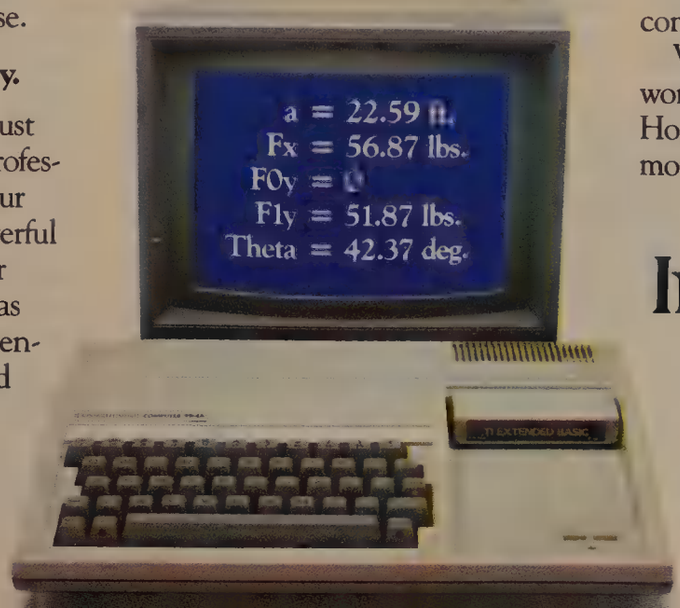
The problem: A 100-foot-long cable, whose weight is 0.67 pounds/foot (corrected for buoyancy) is used to secure a mooring buoy to an anchor in 30 feet of water in a flat-bottomed lake. Determine the horizontal and vertical force components at the anchor and buoy when the buoy is displaced horizontally 92 feet from the anchor. If any of the cable lies on the bottom of the lake, find the distance from the anchor at which it breaks away from the bottom. When the cable is unconstrained (i.e., not touching bottom), the equation of the catenary is $y(x) = (1/k)\cosh[k(x+a)]+b$. See the computer screen below for the solution.

Greater arithmetic accuracy.

Our unique 16-bit processor is just like the one found in expensive professional computers. Coupled with our rich TI BASIC, or still more powerful TI Extended BASIC, it makes our scientific programming capability as good as that provided by many scientific languages for mainframes, and better than some.

We give you 13-digit accuracy, and number range spanning $1x10^{\pm 127}$. All other home computers, using 8-bit processors, limit you to 6 to 9 digits, and only $1x10^{\pm 38}$ range. It's a big difference. And it means you can solve problems on the TI Home Computer which aren't feasible on the others.

TI BASIC also gives you more powerful commands and better flexibility for naming variables, since it recognizes up to 15 character descriptive names. Some others recognize only the first two significant characters in names, which means they can't tell the difference between "TAX" and "TARIFFS." So you have to resort to confusing acronyms.



The TI Home Computer is fluent in seven languages.

You can choose languages for all programming levels, from built-in TI BASIC to TI Extended BASIC, TI LOGO II, UCSD Pascal,* TI PILOT, TI FORTH and our powerful TMS 9900 Assembly Language.

You can also write useful non-scientific programs with the TI Home Computer. It's the only one in its price range that lets you combine 16-color graphics, single-command animation, sound effects, music, and speech with unlimited vocabulary into the same TI BASIC program.

And only TI's home computer comes with a 1-year warranty.

Whatever kind of problem you're working on, work it out with the TI Home Computer. You can't find more computer for the money.

TEXAS INSTRUMENTS

Creating useful products and services for you.

*UCSD Pascal is a trademark of the Regents of the University of California.

Copyright © 1983 Texas Instruments

BOOKS

The earth from space, the most secret agency, dynamical prehistory, snow, fermented foods

by Philip Morrison

MAN ON EARTH: HOW CIVILIZATION AND TECHNOLOGY CHANGED THE FACE OF THE WORLD—A SURVEY FROM SPACE, by Charles Sheffield. Macmillan Publishing Co., Inc. ($29.95). RYDER'S STANDARD GEOGRAPHIC REFERENCE: SATELLITE PHOTO-ATLAS, UNITED STATES OF AMERICA, compiled by Nicholas G. Ryder and Martin Ellison. Ryder Geosystems: Satellite Mapping Division, Suite 304, 445 Union, Lakewood, Colo. 80228 ($75). Every reader knows the Landsat images that so often and aptly illustrate this magazine. Charles Sheffield has for some 15 years been engaged in the transformation and commercial use of satellite images. In his latest compilation he seeks to bring the general reader a thematic sampling of coverage by the first three Landsat birds, reproduced beautifully both for color and for image resolution. If the subtitle promises rather too much for 64 pictures and a brief text, the modest collection is fresh and full of interest all the same.

Two striking spreads take the eye and illuminate the mind. One shows great New York, from Sandy Hook north to Peekskill, and from the Wanaque Reservoir in New Jersey east to Riverhead on Long Island. The second centers on Shanghai as it lies at the silted mouth of the Yangtze. The two spreads appear at about the same scale, although we see a little farther out into the environs of Shanghai. New York is mainly the blue gray of the false-color rendition of built-up urban land, all rough-textured, the strange false-color red of vegetation scattered through the cityscape in parks and edges. Only at the margin do the wooded lands of Westchester and the Ramapos turn the image a clear red. But Shanghai, the home of a comparable population within the scene, is merely an off-center smudge within the broad, flat landscape brilliant with the crimson sign of intense cultivation: rice, cotton, orchards. Everywhere in the Chinese image irrigation canals trace a watery message, as in New York highways inscribe most of the less regular land. Only a couple of new airports stand out as linear disturbances in the thorough Chinese cultivation.

The scanner shows the Dutch in the act of making the Netherlands: their polders grow out to reach the long dikes that turned the Zuider Zee into a steadily diminishing lake. The size of the task is impressively communicated on these pages; all but a few of the images have roughly the same scale, some 10 or 12 miles per inch of image, so that the eye learns. The pictures show several well-known dams, such as the Nile dam at Aswan; the lesson here is clear as well. Any dam (the best example shown is the new La Grande Hydro-electric Complex near James Bay in central Quebec) is small and hard to find in such images; its mass is hardly to be noted in the landscape, but the tiny block of concrete has all the same cleverly ponded up a very large pool of water.

Chicago is neatly imaged, the mile-wide ring of the Fermilab accelerator plainly visible 30 miles almost due west of the Loop. Paris is viewed in the spring, and Cairo. The pyramids do not show, although they were on camera; the minimal picture element is about an acre in area, and a patch of a few acres is hard to see unless there is some special contrast. The Valles Caldera above Los Alamos is one striking place, and the desert site of the first test of an atomic bomb is also offered, along with a number of pictures evoking war under the rubric of strategic pressure points. The Empty Quarter is caught in one shot, inhabited by no life at all save that of a single djinn, whose tower of smoke by day extends 10 or 15 miles downwind. He appears in the form of a great gas flare among the dunes, and there is his road of entry from Tarif on the Arabian Gulf.

Man on Earth is a visually attractive and interesting sampling of city and wilderness in change, of farming and dwelling places, of trade and even of the past, all from 500 miles aloft. The text pages are readable, one page to an image along with some brief introductions to the categories, and the key maps, from Vatican City to K-2, are helpful.

Satellite Photo-Atlas, United States of America is a more austere gift out of the same treasure. In 1974 the Soil Conservation Service assembled a mosaic in black and white of single-channel scans from Landsat 1 showing the U.S. as a whole. The heroic product, consisting of more than 600 prints, occupies an entire wall; a page-size reproduction is inexpensive but lacks useful detail. Some years ago this department reviewed a book of reproductions of that photographic map, made up into an inexpensive U.S. atlas; this is a second publisher's compilation from the same source. It is not cheap (its main users will be professionals, say pilots and planners eager to order just the right Landsat prints for themselves), but it is very well printed, much better than the earlier version was.

The book has an exceptional apparatus, including a red overlay of main borders, highways and cities, a mark for every air navigation (VOR) beacon, which is a fine means of careful comparison with other maps, a latitude and longitude grid and a full gazetteer. It presents a real orbital atlas of the 48 contiguous states over about 80 spreads with a useful overlap, at a uniform scale of one to a million. (Hawaii is here at the common scale, but not of course in correct position, and big Alaska is at a much reduced scale.) The one weakness is made clear by looking for Ithaca, N.Y., which lies all but lost in the gutter of a spread. The compilers plainly dream of a reliable atlas of the world at one-millionth scale, a goal long held by cartographers and never yet accomplished.

THE PUZZLE PALACE: A REPORT ON AMERICA'S MOST SECRET AGENCY, by James Bamford. Houghton Mifflin Company ($16.95). The National Security Agency is better known nowadays than, say, the National Reconnaissance Office, its spy friend in the sky. Most of what is publicly known of the shadowy NSA is compiled or cited here in a responsible investigative study of the agency, from its roots in the black and purple annals of U.S. cryptology since World War I through its official birth in 1952, by Presidential charter of seven still secret pages, to its present powers and aspirations. Its director, always an able general or admiral (the incumbent is the tenth), was the "most influential individual" in the U.S. intelligence community, in the opinion of the Senate Committee on Intelligence in 1976.

That status has hardly diminished; an estimate of the employee strength of the NSA comes to some 68,000 people, two-thirds of them perhaps in the big headquarters complex at Fort Meade (with 10 acres of computers alone, they say) and its annex adjoining the Baltimore Airport. Others serve the NSA around the world; within the three uniformed services another considerable force carries out related tasks usually at a more tactical level, not counted in the NSA strength. The headline-shy NSA is likely

an order of magnitude larger than its celebrated cousin the Central Intelligence Agency. Every day the NSA tries to dispose of 40 tons of classified waste, and in 1980 it generated new classified documents at the rate of one or two million per week, more than all the other components and agencies of the Government combined.

Its job is eavesdropping. Puzzle Palace naturally speaks all the tongues of Babel. Its Crays and its gifted mathematicians, both in-house and at its think tank in Princeton, preserve the high tradition of code breaking from the old glories of Magic and Ultra. Today, however, the hero is the communications engineer and the quarry is the river of bits that fly around the world; the sport is rather like commercial fishing. Most codes are either unbreakable or easy, a matter for budgeting. A far-flung net of dishes samples the stream, the computers set to detect the telltale headers of the address and the names on the watch list. Some 15,000 half-circuits carry almost all the commercial telephone and data traffic overseas from the U.S. There are 75 million foreign telegrams and telex messages each year and plenty more in voice, to say nothing of the data channels. Four U.S. earth stations, two east and two west, handle 10,000 circuits to and from the synchronous satellites that girdle the earth.

Bamford makes a strong if circumstantial case that the NSA has placed big dishes near each of these terminals, right in the footprint, sampling what they will of the American bit stream. About a third of the circuits overseas still travel on the Transatlantic Telephone Cable. Ninety percent of that cable traffic comes ashore in Rhode Island, securely vaulted, hard to tap. At the cablehead, however, those same signals are beamed by microwave to relay points in Connecticut for the domestic network of long lines. For that small hop they are wide open to unobtrusive horns of the NSA. Entirely domestic traffic is safer so far, protected by its daunting volume, as so far real voice messages, even pulse-coded, are safer than fully digital traffic. The diplomatic pathway from Washington toward New York enjoys no such anonymity; NSA towers stand, perhaps by chance, right along the microwave beam.

The battle is not always to the strong. The encryption of private computer data by a clever IBM chip established a national algorithm, certified as workable by the NSA. Since 1977 it has been the civilian standard. Its days are numbered as brute-force search gets cheaper. It was argued that a dedicated computer with a million custom chips monkey-searching at one microsecond per permutation could break that code in half a day: no big drain on the NSA budget. There are informed if cynical



Another QUESTAR® . . .totally new. . . the long-range microscope

This is the QM1, the new instrument by Questar Corporation. It focuses at distances from 22 to 77 inches from the target. With it you will inspect hot materials at safe distances, observe live specimens in natural habitat, inspect micro-circuits, use it for crystal growth studies and a thousand other applications. It will let you scrutinize and photograph living organisms or radio-active samples with no damage to either one or yourself. And the magnificent design of these optics permits resolution below 2.5 microns. With a selection of eyepieces and extensions, magnifications can range as high as 150X.

Shown below on the right is a chip from a micro-circuit photographed with a normal 50mm lens to show its tiny size. On the left is a QM1 photograph of the lower right corner, reproduced from an 8 × 10 enlargement of 35mm Kodak Technical Pan. Distance 22 inches.

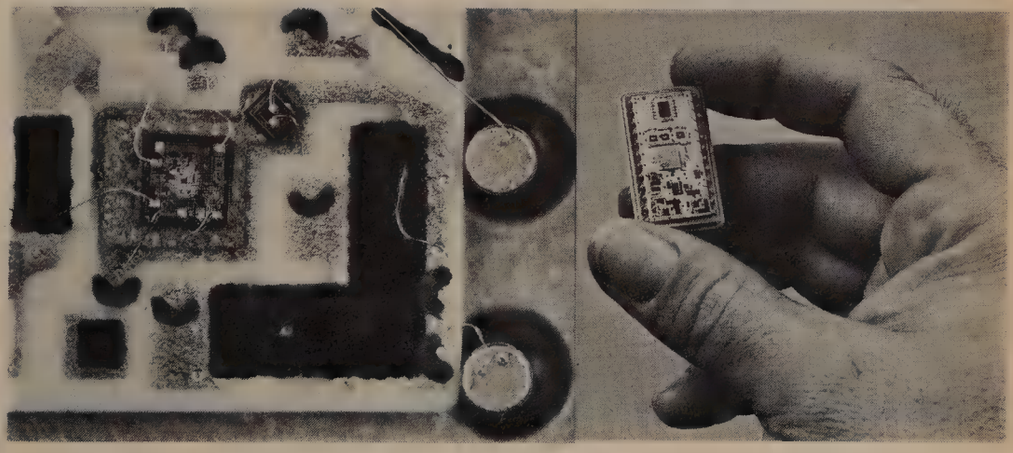
With the Questar QM1 we can provide full support systems of equipment and accessories, including video and high-speed motion cameras, lighting equipment, micro measurement and image analysis, and computer storage and display units.

• • •

Questar Corporation, one of the recipients of the 1983 IR100 awards, was cited by Industrial *Research and Development for creating one of the 100 most significant technical products of the year. This climaxes for us the enthusiastic reception of the QM1 by scientific laboratories and by industry, wherever there is a a need to observe or inspect minute elements in non-destructive and non-disruptive situations.*

QUESTAR

P.O. Box C, Dept. 205, New Hope, PA 18938
(215) 862-5277



critics who feel that the NSA review was less than disinterested. If the original IBM key, 128 bits long instead of the mere 56 bits adopted, had been accepted, even the powerful NSA would have had no chance to read the coded bits. The conflict over the public-key codes during the past five years is a sign of NSA concern over the spread of cryptologic prowess outside its control. It is unexpected that in this thick organizational history and anatomy there is hardly another line about today's code-breaking techniques.

Satellites orbit high and low, planes fly, ships sail, army posts are alert, all ferrets for microwave signals, telemetry, routine radio traffic from planes and ships and vans, whatever the widespread net can catch. Back it comes, now translated into the wide-band channels of the special intelligence communications network, mainly satellite-borne. Down it goes to the dishes waiting at Meade and then inside to the thousands of analysts, linguists and code breakers. Maybe all this is effort well spent for a great power. The flashy successes we hear about in due time do not make an open-and-shut case. The Pentagon Papers threw some light in these dark corners. An Army Security Agency group in our Moscow embassy had found how to intercept and to unscramble casual local radio-telephone conversations from the limousines of such worthies as Leonid Brezhnev and Alexei Kosygin. Imagine the prize intercepts, duly encoded and proudly sent back stamped TOP SECRET VIPAR GAMMA GUPY. "It was very gossipy," one official remembered, although he thought it had been extremely valuable on the health and personality of the top Russian leaders, one of the best pipelines we had! Yet they never exchanged a word about, say, Czechoslovakia.

Naturally we would prefer not to be done by as we scheme to do unto others. The NSA works hard to preserve communications secrecy on our part, by discipline, ciphers and ingenious hardware. There are three well-known ways to hide a radio signal from interception; they are clever. Two approaches rely on time division. A wary submarine surfaces to signal home. It emits its entire message in one split-second burst. It may reasonably hope the frequency channel chosen will be free from surveillance for that brief moment. Another device divides a message from a clandestine agent hiding within some "denied area" into millisecond segments, to bounce them out one after another on many different frequencies, chosen in a pseudorandom order known only to the prepared listener, the signal unheard within the radio hubbub of a big city. A still subtler scheme dilutes the signal with a great deal of noise and sends it all on a wide band. A clever counter to every such clever game in the domain of frequency and time is to be expected.

The sources for this book are themselves an example of intelligence gathering. Hearings in Congress are obvious and major. Interviews with old-timers once in high places are full of interest, and they may or may not be relied on. Visits and telephone calls help, even when the unwilling recipients are suspicious and guarded. What about the archived celebratory papers of leaders long gone, the personnel newsletters hard for any large bureaucracy to avoid, the accounts of the Credit Union of Federal employees, the occasional defector now a refugee in the U.S.S.R., technical manuals for military personnel? Then there are the counterparts in Britain, closer kin than the CIA. They are all made use of here, in a first approximation to an unfolding of this complex and still hidden history.

The future? Fiber optics would seem to tax the Big Listener, whereas packet-switching networks smooth the eavesdropper's way. Whisper privately behind your hand if you will, but it is certainly optimistic to expect any signal cast afar to remain long exempt from so exigent a busybody.

IN PURSUIT OF THE PAST: DECODING THE ARCHAEOLOGICAL RECORD, by Lewis R. Binford, with the editorial collaboration of John F. Cherry and Robin Torrence. Thames and Hudson Inc. ($18.50). "I'm the oldest new archaeologist in town," our author remarks. The phrase does no bad job of conveying the tone of this book, a book for the general reader as deep as it is breezy, compiled mainly from a group of loosely connected lectures Professor Binford gave to British audiences a couple of years ago, a good way from his own Albuquerque campus. His message is productively iconoclastic; the images he shatters, sometimes rudely, have mostly lost their magic already. He is no cynic but a man with hopes for a rational future, when better methods will allow us knowledge that grows, not merely forever-shifting intellectual fashions as transient answers to questions as difficult as "Why did it happen?"

The dynamical past is beyond our reach. What we can know is a preserved static record. We can, however, learn to read that record. We do not read it right by quick conclusions from empathy and implicit assumption. The first example pleasingly outlined here is the way early man became the bloodthirsty killer, to fit Raymond Dart's own observations of bone frequencies in the South African caves that did not fit the skeletal facts. The selected bones were brought home by those ruthless old boys, Dart held. Then C. K. Brain's brilliant studies made clear that the collection had been assembled by cats; the australopithecines had not lived in the caves but had merely been eaten there!

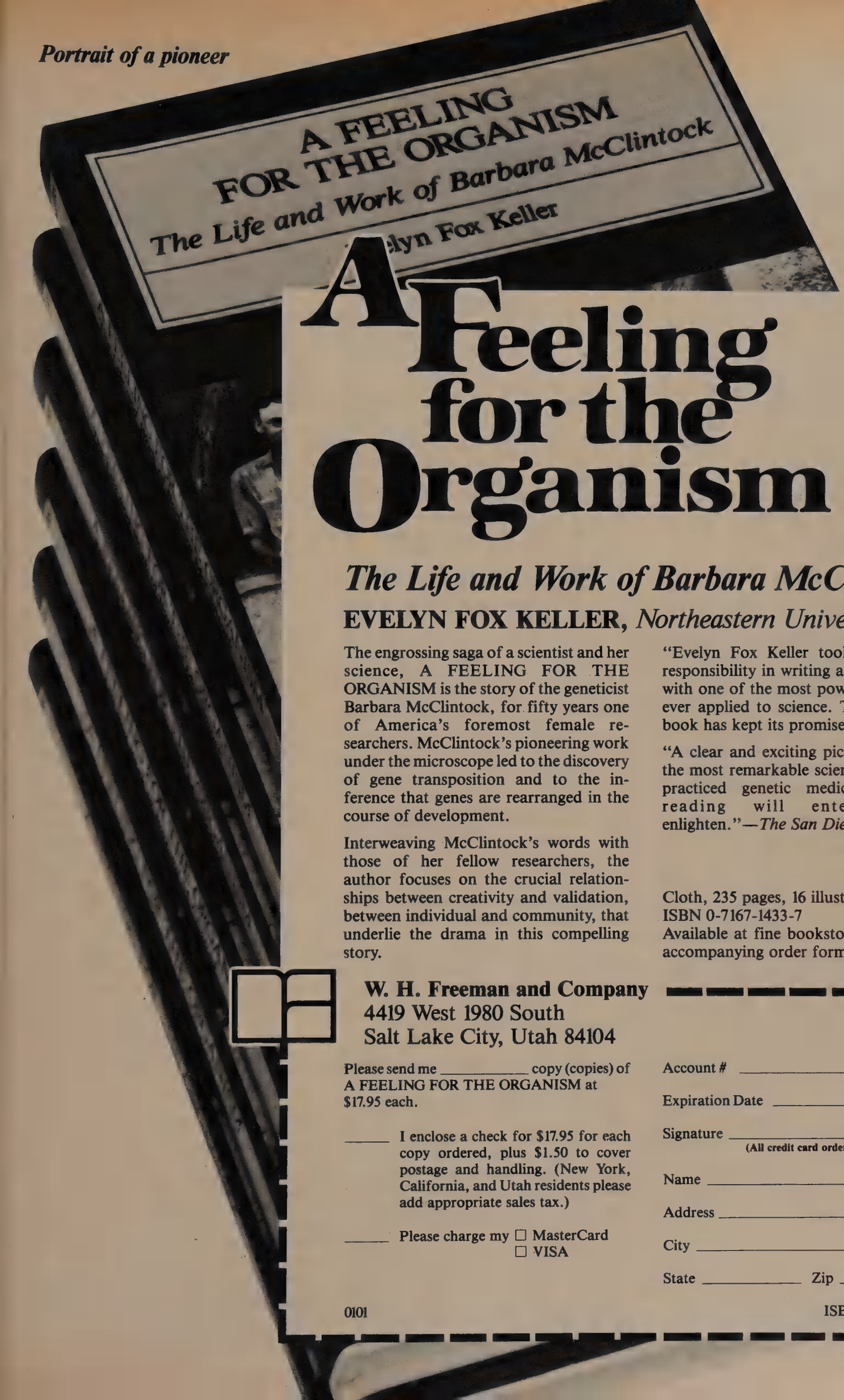
The new vision from Olduvai and the rest of the Rift sites again casts man as the mighty hunter, although now no cannibal. His killing floors yield ancient tools along with plenty of animal bones. Hasty orthodoxy sees the sites as dwelling spots, home bases. In actuality heaps of bones are not found today anywhere so much as at water holes, where they accumulate from kills by predators, just as much with hominids absent. The oldest Rift deposits hold the hammering tools; the cutting flakes lie later, high up in the slowly deposited strata. A new scenario enters: the hominids began by breaking bones for marrow, a minor dietary item; only later could they compete for meat, once flake tools could cut like a hyena's tooth. All that time our forebears were no mighty hunters but "the most marginal of scavengers" on the kills of the fiercer cats. The point, however, is not this succession of more and more plausible inferences; it is the need for the elaboration of detailed studies of the old sites, and in the present for studies wherever analogies exist, until we can stipulate some test of our exciting inferences.

The longest section of the book is an account of Binford's field work among hunters today in the tundra and the desert. It is detail he seeks among them, significant detail, adding up in the end to an entire system of life. The Nunamiut do not live and hunt in one way; they live by many. They build now rings of stone for tents and again set hunting stands with hearths, simple for one hunter, complex for many. Around the central hearth there is a ring of tossed debris, should many hunters encircle the fire; if only one hunter is present, the toss pattern is quite different. Tented and open-air sites leave distinct toss patterns.

By now most scholars agree that Binford and not his late friend, François Bordes, was correct in maintaining that the four distinct stone-tool kits found at Mousterian (Neanderthal) sites marked not distinct subcultures of the people but one and the same hunters turned to different functional tasks. The time span within the finds is very long, and people in their times and their life spaces arrange themselves in many ways. Ethnoarchaeology, experimental site-making and the quantitative study of the documents (particularly photographs) from the 19th century, when hunting peoples were more abundant, are three areas only now beginning to receive the attention they demand, if ever we are to draw firmer inferences from our digs.

The origins of agriculture and even of complex social order itself remain as the grandest questions since the time when human primates were coming to be differentiated from their cousins. Here there is more to criticize than to

Portrait of a pioneer



A Feeling for the Organism

The Life and Work of Barbara McClintock

EVELYN FOX KELLER, *Northeastern University*

The engrossing saga of a scientist and her science, A FEELING FOR THE ORGANISM is the story of the geneticist Barbara McClintock, for fifty years one of America's foremost female researchers. McClintock's pioneering work under the microscope led to the discovery of gene transposition and to the inference that genes are rearranged in the course of development.

Interweaving McClintock's words with those of her fellow researchers, the author focuses on the crucial relationships between creativity and validation, between individual and community, that underlie the drama in this compelling story.

"Evelyn Fox Keller took on a heavy responsibility in writing about a woman with one of the most powerful intellects ever applied to science. This ambitious book has kept its promise."—*Discover*

"A clear and exciting picture of one of the most remarkable scientists who ever practiced genetic medicine, and its reading will entertain and enlighten."—*The San Diego Union*

Cloth, 235 pages, 16 illustrations. $17.95
ISBN 0-7167-1433-7
Available at fine bookstores, or use the accompanying order form.

W. H. Freeman and Company
4419 West 1980 South
Salt Lake City, Utah 84104

Please send me _____ copy (copies) of A FEELING FOR THE ORGANISM at $17.95 each.

_____ I enclose a check for $17.95 for each copy ordered, plus $1.50 to cover postage and handling. (New York, California, and Utah residents please add appropriate sales tax.)

_____ Please charge my ☐ MasterCard ☐ VISA

Account # _____

Expiration Date _____

Signature _____
(All credit card orders must be signed.)

Name _____

Address _____

City _____

State _____ Zip _____

0101

ISBN 0-7167-1433-7

propose, and Binford cuts away sharply, mainly at undemonstrated preconceptions of causality, from aristocratic trade to sedentarism. If we cannot yet clearly infer from a pile of bones whether our ancestors were there the hunters or the hunted, we may have little chance to settle higher issues. "We cannot just follow the Muse of Curiosity and speculate." Certainly we cannot fall back "to that masquerade of productive thought currently parading under the rubric of sociobiology."

One worry remains when the reader comes to the last of the splendidly combative footnotes: it may be that there is an order of magnitude lacking in the overall richness of archaeological investigation, a poverty of sites, models, techniques and personnel that must be filled in before we can expect the bite of inference. Ought we to hope to understand how and why agriculture appeared more easily than we learn the causes of, say, the French Revolution? Enough has been said in review; the short, vivid book will please, prickle and enlighten almost any reader.

HANDBOOK OF SNOW: PRINCIPLES, Processes, Management & Use, edited by D. M. Gray and D. H. Male. Pergamon Press, Inc. ($60; softcover, $20). Canadian readers will very likely see this season's snow arrive before the end of November, and most of the rest of us North Americans can expect it a month later. So the maps suggest, as they exempt the dwellers in California valleys. Snow cover is pretty much a phenomenon of the Northern Hemisphere; there is God's plenty of snow cover in the Antarctic, of course, but the other three big southern land masses have almost none except in the high mountains. This thick volume is the work of two dozen Canadian scientists and engineers (one contributor comes from the somewhat more southerly latitude of Hanover, N.H.), who pool their learning and experience into a comprehensive reference text that regards snow not only as a phenomenon but as a resource and challenge. The 20 chapters fall into four parts: a survey of the importance of snow to living forms, particularly ourselves and our domesticates; a thorough summary of the formation of snow and its physical nature, of how we measure it and of its motions fast and slow; an engineering view of how to travel on snow, how to scrape, blow and melt it, how to build against it and how to control drift, and two closing chapters on the mechanics of skis and skiing.

Snowy mountains lie at the headwaters of arid river valleys in many lands. The Yellow River, the Ganges, the Indus and the Rio Grande are four snow-fed rivers; the irrigation waters they provide supplement local rainfall to grow the grain that nourishes one human being in five or six. The estimate of snow cover and its rate of melting are plainly a precondition, still not very well met, for a rational agriculture worldwide. Radar and satellite images can measure snow-cover area, but snow is a complex blend of water and air; its yield on melting is rather more hidden. There are 3,000 marked traverses in the U.S. and Canada along which cores are taken by hand at five or 10 marked points several times each winter to give runoff estimates. Recently airborne detection of the flux of natural gamma rays from radioactive potassium and uranium in the soil has been tried; a traverse in front of the snow calibrates the attenuating effect of the layer of water.

None of this can be precise or easy; even in the laboratory and the wind tunnel, let alone the frozen tundra or the alpine slopes, "there is no material of engineering significance that displays the bewildering complexities of snow." Each snowflake begins as a single crystal, growing out of a supercooled water droplet nucleated usually on an airborne particle of clay silicate. It starts as a simple form, often merely a tiny hexagonal plate of ice, less than a tenth of a millimeter across, slowly falling. Soon the crystal grows by vapor deposition from the water droplets around it, whose vapor pressure exceeds that of the ice. Its habit depends on the ambient temperature: the relative growth rate of different crystal faces varies. The competition is lively enough to allow all kinds of forms, from needles to hollow tubes to the six-sided mineral flowers that delighted old Kepler. Now the bigger crystals fall faster; they collide and accrete, if for no reason other than that their wake in falling attracts neighbors 40 diameters behind. Snowflakes, certainly a ripe subject for fractal studies, show their intricacy in the empirical relations for mass, which may increase only linearly with diameter.

Now the work of the engineer begins. (Efforts to nucleate snowfall aloft have succeeded, and a decade ago a University of Washington snow-making group even called their shot on a predesignated target area in the Cascades.) A layer has formed of heterogeneous density, strength, heat conductivity, hardness, surface friction and albedo. (Snow is, incidentally, quite black for radiation near a wavelength of 1,500 angstroms in the ultraviolet.) The fallen snow, never very far below its melting temperature, is not stable; it tends to lose free energy by sintering and by sublimation, its internal surface diminishing as small crystals give way to larger ones, and by settling under its own weight. Lying under stress, as on a slope, the snow emits sound at subaudible frequency, probably from the continuing fracture of stressed grain boundaries. Once these accumulate enough to join into a continuous fracture, the snow slab gives way under gravity. Avalanche! Recently there is direct evidence for this view: noise detected before failure.

The fantasies of popular building in Los Angeles owe a good deal to the absence of snow load. Car washes in flimsy Gothic are not open to Canadian architects. Discipline is enforced by roof failure. Provision is made to support 80 percent of the ground snow load in average exposures, which may amount to 40 percent of an atmosphere in snowy Quebec. Snow is transported by the wind, of course; most such transport, like that of desert sand, takes place within an inch or so below the surface (a photograph documents the motions), but slat fences reduce the speed of wind transport, so that drifts form both to lee and to windward. This useful control over drift has its convincing opposite in the blower fence, like "inclined table tops mounted on posts." Such structures, with the low edge downwind, act to speed up the wind and to erode the snow to a distance of almost 20 feet, moving the local drift away from any obstruction surmounted by the fence.

Toronto is well served by its snow-removal system. Continuous mechanical plowing, with salt and sand as needed, serves to keep snow layers on the expressways below a depth of an inch for 98 percent of the storms, even those with a snowfall rate of up to nearly an inch an hour. The formula for the number of blade-equipped trucks and sanders per kilometer of highway is offered for a wide range of service quality. The cost is not small. This winter is the 100th anniversary of the big rotary snowblowers by road and rail, the winter glory of the old newsreels. Their design theory is given here, although it is not clear that Orange Jull of Orangeville, Ont., the mechanically proficient miller who designed the first one, needed it.

The two million snowmobiles of this continent, the noisy and effective vehicles that carry their drivers out into once-silent snowy lands mainly for sport, trace their ancestry to the first Ski-Doos, produced in the winter of 1959–60 by Armand Bombardier of Valcourt, Quebec. He began work on an over-snow vehicle in 1922 at the age of 15, with a Model T and parts of a sleigh. The standard snowmobile has a single wide rear track and a steering ski or two in front. They can race at a mile a minute and climb 45-degree slopes. The Eskimo hunters use them, but the fuel-hungry vehicle cannot provide the warmth of sled dogs for survival for hours under the most trying conditions of Arctic travel. Heavy burdens can be carried across snow by a wonderful variety of buses and trucks; the largest designs call for huge, ribbed pneumatic tires at very low pressure, some made easily adjustable to the needs of the ter-

rain. A man walks with a bearing pressure of about a third of an atmosphere; the big soft-tired trucks can follow him pretty easily across a four-foot snowbank with a load of 30 tons. Snowshoes and skis, however, reduce the bearing-pressure requirement by a factor of 10.

The volume is full of interest, a bargain in softcover for any technical reader inclined to the "enhancement of life in a snow environment." It is a model effort to cope as rationally as can be with the genuine complexities of a real world, from the layer of disordered protons that lies at the frictional surface of snow to the tax burden of busy, snowy Ontario.

HANDBOOK OF INDIGENOUS FERMENTED FOODS, edited by Keith H. Steinkraus, editor in chief; Roger E. Cullen, Carl S. Pederson and Lois F. Nellis, associate editors, and Ben K. Gavitt, assistant editor. Marcel Dekker, Inc. ($79.50). The sweet life may benignly turn a little acid, as for one who enjoys a slice of Cheddar on a fragment of the sourdough loaf dear to that luxurious city along the Bay of San Francisco. This particular bread, prized for a moist, chewy texture and the faint sharpness coming from fermentation to lactic acid, is made to rise by a leaven of wheat flour and water kept wet and warm until it begins to ferment. A novel yeast, not at all the usual bakers' yeast, and a number of specific lactose-fermenting bacilli form the essential flora of the sourdough. The organisms nourish each other in a balanced way: the bacteria metabolize maltose to yield glucose, which the yeast requires, while the yeast secretes the amylases that catalyze the change of wheat starch into maltose. The trick for maintenance is frequent subculturing into fresh dough, part of the time-honored process. San Francisco, although a fine place for the flora, is not unique. The very same microecology has been demonstrated in sourdough leavens both in Glasgow and in Poland. The substrate selects for the well-fitted organisms, airborne far and wide in the temperate world.

Here is a thick and authoritative volume in which the sourdough phenomenon makes only one rather provincial example in a catalogue raisonné of fermented foods for the widely cosmopolitan palate. There are about 250 substrates discussed, from almonds to sea cucumbers, from the durian fruit to the winged bean. Even wastepaper nourishes something good to eat: the oyster mushroom. For most of the hundreds of products there is an account of importance and use, a recipe in the form of a flow chart, an introduction to the microbiology and the biochemistry of the process and its results, and an estimate of nutritional changes and values. The book fed on papers delivered late in 1977 by a couple of hundred experts from around the world, assembled in Bangkok to outline the indigenous fermented foods they knew. That rich broth was refined, compressed by a factor of about four, and is presented with something of a flourish by the corps of expert editors, who work in the food sciences at Cornell University.

Few will claim to have appraised the relative qualities of ambali and bongkrek, ginger beer and jackfruit wine, munkoyo and ragi-tapai, trahanas and urwaga. Our world is large, even for the jettiest set. All of those are important items of food or drink somewhere, usually prized by the poor for their good value and favored by the discriminating for variety and pleasure. It should be pointed out that most cheeses made from milk and the hops-flavored malted-barley beers of Europe, now enjoyed the world around, are not systematically included.

Main divisions treat the dishes in their imposing variety in four distinct classes. There is hardly space to summarize even one example of each class to characterize the generous global buffet. In the class of acid fermentations used to preserve and to enhance the qualities of fresh foods there is the section on acid-leavened breads and pancakes, about 60 pages of matter from a dozen countries. Sourdough bread is a simple yeoman of this family. Perhaps the subtlest is idli and its group of kindred foods. These are staples in southern India and in Sri Lanka. Two slurries are carefully made, one of soaked and coarse-ground rice, the other of dehulled and fine-ground black gram, the commonest bean of the region. They are mixed into a thick batter, salted lightly and kept overnight in a warm place. No inoculum is needed; the key organisms, two bacterial forms, are present and outgrow all competitors to produce both carbon dioxide as leaven and a pleasant acid flavor.

The idli are made up into little white cakes and steamed in a special pan. The same batter, a little thinner, makes dosa, a lunch or dinner dish of large crisply fried pancakes. There are many other variants, and a range of traditional hot condiments and accompaniments of virtue. Early work appeared to show that the limiting amino acid of the bean protein was increased in amount, much improving the nutritive quality. This has not been confirmed; it is clear, however, that the vitamin content is enriched. Like fresh beer in the medieval European winter, idli are good for you.

Take another class, the foods developed as protein-rich vegetarian meat substitutes, with a fibrous meatlike texture. A large family of Indonesian preparations, generically known as tempe, fills this part of the volume. Two fermentations are involved. The first is a bacterial acidification; it arises naturally when dehulled soybeans are soaked in water. It is not well understood; in the laboratory adding a little acid during soaking is a replacement. The second stage is grander: a particular mold species grows on the partly cooked and long-soaked beans, but it must be supplied as an inoculum. In Indonesia the mold is sold grown and air-dried on the leaf of a hibiscus; in Tennessee, say, a modern small factory industry supplies the growing American market among vegetarians for tempe, providing spores and soybeans in less romantic packages.

The mold can be any of a number of species of a couple of genera. The mold spores are mixed with the drained beans and incubated for a day or so at blood heat. All these conditions are flexible; the process has been studied in detail for many decades. The mold overgrows the beans completely, knitting them together with the tender fibers of its mycelium. The cake, once overgrown, can be sliced and fried in a few minutes; it takes hours to cook untreated soybeans, and then they lack the nutlike flavor and meaty texture of tempe, which is also a little peppery. Best of all, the process can be used for many kinds of substrates besides soybean: peanut presscake, coconut-milk residues, other beans, peas, even wheat and barley.

A start has been made on the engineering of such solid-state fermentations, with pumped air and plenty of controls. The authors tout the process as one way of making substitutes for animal protein in the world of 10 or 12 billion diners to come: inexpensive, easily digested and good to eat. (This is unconfirmed here; the reviewer has yet to try his first tempe-burger.) Plainly, eating a mold colony entire is ecologically more efficient than dining on a soy-meal-fed chicken encumbered with feet, feathers, bones and beak, let alone a horned beast of the field.

The volume is completed by soy sauce and sake, fully industrialized in Japan; sauerkraut and miso; a wide variety of tasty-sounding alcoholic brews from homes around the world, based on honey, sorghum, sugarcane, palm-tree sap, bananas, agave or cornstalk juice; together with a set of complicated pastes and sauces compounded of salted fish and shrimp (including a few flow charts that deeply challenge the cultural conventions of any American). Several interesting essays, pointing out relations between indigenous technology and its adaptation to both large scale and new places, come at the end. There are plenty of problems: sauerkraut is sometimes a surprising pink (a pigmented yeast), some false-banana sourdoughs can induce abortion and there are well-known fatal mycotoxins to be shunned. The book is no ordinary compilation but a powerful eye-opener to any self-centered view of everyday life.

Established 1845 SCIENTIFIC AMERICAN November 1983 Volume 249 Number 5

The Uncertainties of a Preemptive Nuclear Attack

It is said that the U.S. needs new land-based missiles because its present ones are vulnerable to attack. Analysis of uncertainties in such an attack suggests that the vulnerability is exaggerated

by Matthew Bunn and Kosta Tsipis

The deterrence of nuclear war requires that nuclear forces not be vulnerable to a preemptive attack. For decades strategic planners in both the U.S. and the U.S.S.R. have been concerned about the possibility of a disarming nuclear first strike, which would leave the victim of the strike unable to retaliate in kind. Each nation has expended large resources on efforts to maintain the survivability of its strategic nuclear forces.

Concerns over the vulnerability of land-based nuclear forces were heightened by the development of multiple independently targetable reentry vehicles (MIRV's) in the late 1960's. This technology, tested initially by the U.S., makes it possible for one missile to carry several warheads, each warhead capable of striking a separate target. Thus each MIRVed missile might be able to strike several of the opponent's missiles, each of which might itself contain several warheads, giving the attacker a double advantage.

As a result the evolution of Russian MIRV technology generated fears that once the U.S.S.R. had developed MIRV's with the appropriate combination of accuracy, yield and reliability it would have the ability to destroy most of the U.S. land-based missile force in a first strike, using a fraction of its available weapons. Although land-based missiles constitute only a fourth of the U.S. strategic nuclear forces, with the rest represented by submarine-launched ballistic missiles (SLBM's) and bombers, concern has arisen that the vulnerability of the ICBM component of the strategic "triad" alone would present the U.S. with a grave problem of national security.

In the fall of 1977 the worst fears of U.S. strategic planners were realized. The U.S.S.R. began a series of tests of a new guidance system with greatly improved accuracy. Simplified calculations indicated that once the U.S.S.R. had deployed an adequate number of these more accurate MIRV's it would indeed have the ability to destroy the bulk of the U.S. land-based missile force in a first strike. Thus was born what has come to be called "the window of vulnerability." This concept has dominated American strategic thinking for some years, providing the primary justification for the development of a new generation of U.S. strategic weapons, including the MX and Trident II missiles.

Simplified calculations do not, however, do justice to the substantial uncertainties inherent in any assessment of the results of a countersilo attack. Intelligence information is rarely absolute, and when strategic planners are confronted with uncertainty about the actual value of such parameters as the accuracy of the other side's ICBM's, they must make assumptions that are conservative from the defender's point of view. Unfortunately the process often obscures the fact that any attack would also involve substantial uncertainty from the attacker's point of view.

Because of the immense destructive power of modern nuclear arsenals, any nuclear first strike would represent a gamble on a scale absolutely unprecedented in human history; the future of entire civilizations would hang in the balance. As a result of the magnitude of the stakes any uncertainty about the outcome of such an attack will act as a powerful deterrent; such gambles are not taken without extremely high confidence in the outcome. The question for anyone considering how to plan the attack will always be not only "What is the expected outcome?" but also "What is the worst plausible outcome?" Therefore in assessing the possibility of an attack it is crucial to gauge quantitatively the uncertainties involved, in order to develop an assessment that is "attack-conservative." In U.S. assessments of the strategic balance this is rarely done. What is generally presented to Congress and the public are the results of an idealized, nearly flawless attack; the uncertainties inherent in the attack are often ignored. In this article we shall be offering a corrective.

The silos in which modern ICBM's are housed are underground concrete structures, "hardened" to withstand the effects of nuclear blasts. There is a wide range of nuclear effects that might damage an ICBM within such a silo, but spokesmen for the U.S. Air Force have indicated that U.S. ICBM silos are generally most vulnerable to the shock wave of the nuclear blast. Thus the hardness of a given silo is usually expressed in terms of the shock-wave overpressure required to destroy it, measured in pounds per square inch (p.s.i.).

The silos currently housing the 1,000 U.S. Minuteman ICBM's are generally estimated to be capable of withstanding overpressures of up to 2,000 p.s.i. The detonation of a half-megaton weapon, such as those carried by the most accu-

rate Russian MIRV's, would create such overpressures at ranges of roughly 300 meters; therefore to destroy a Minuteman ICBM a half-megaton weapon would have to be detonated within 300 meters of the silo. Thus the accuracy with which the weapon is delivered, although it is not particularly important in an attack on a city, would be of decisive importance in any attack on hardened targets such as ICBM silos.

The ICBM's with which such weapons are delivered have three major parts: a rocket, which may have several separate stages; the payload, which consists of one or more reentry vehicles (RV's) armed with thermonuclear warheads, and the guidance system, which directs the rocket thrust in order to place each RV on the appropriate trajectory for it to reach its intended target. When such a missile is launched, the main rocket fires for only the first three to five minutes of the flight. Over the next few minutes a smaller rocket known as the postboost vehicle provides the final trajectory adjustments necessary to set each RV on the path to its intended target. The RV's are then released and fall freely in the earth's gravity field, unpowered and unguided. At the end of a half-hour flight the RV's reenter the atmosphere and detonate over their targets. The missile's flight can therefore be divided into three phases: the boost phase, free flight and reentry.

To deliver the RV's within several hundred meters of their targets over ranges of 10,000 kilometers calls for a highly sophisticated guidance system. Current strategic weapons rely on the technique known as inertial guidance, in which gyroscopes and accelerometers are used to measure the specific forces acting on the missile. These inertial instruments cannot, however, measure the force of gravity, because of the equivalence of gravitation and acceleration described by Albert Einstein: an accelerometer in a free-falling elevator would register zero, even though the elevator would be accelerating toward the ground under the influence of gravity. Therefore in order to account for the effects of the earth's gravitational pull on the motion of the missile a mathematical model of the gravity field as a function of position must be programmed into the missile's guidance system prior to launching. By combining the gravity model with the measurements of the specific forces made by the accelerometers, the guidance computer can use Newton's laws of inertia to calculate the motion of the missile in three dimensions and put it on the appropriate trajectory to its target.

The accuracy of such a weapon can be affected by errors from many sources. The two main sources are errors in the inertial-guidance system and errors associated with reentry into the atmosphere; in some current ICBM's the deviation from the target resulting from either of these sources is 100 meters or more. The errors in the guidance system originate with imperfections in the gyroscopes or the accelerometers; they can be either constant, accumulating with time, or random, arising from vibration, shock or changes in acceleration. The errors associated with reentry originate with unpredictable atmospheric variations over the target, such as winds and variations in atmospheric density, and with uncertainties in the ablation of the RV's nose cone. Because the RV from an ICBM reenters the atmosphere at a speed of some 7,000 meters per second, the nose cone will begin to burn away as the RV passes through the atmosphere; this ablation tends to occur in a rather unpredictable and asymmetric manner, giving rise to aerodynamic instabilities that degrade the accuracy of the RV. Moreover, severe weather effects, such as heavy rain or snow, can cause drastic increases in the rate of ablation, further reducing the accuracy of the RV.

Lesser sources of error include errors in the gravitational model relied on by the inertial-guidance system, errors in warhead fusing and errors in the determination of the position of the target. Each type of error can contribute several tens of meters to the system's overall "error budget." The determination of the initial position and velocity of the launcher is extremely important in systems fired from mobile platforms, such as SLBM's, but is negligible for silo-based ICBM's; indeed, this type of error is one of the reasons current SLBM's are considered to be too inaccurate to strike at hardened missile silos.

The bulk of the error sources of a ballistic missile will be random from one missile to the next. As a result if a large number of such weapons were fired at a single target, they would tend to fall in a random scatter around the target. The spread of the scatter is measured by the parameter known as the cir-



COUNTERFORCE TARGETING PATTERN is suggested by this photograph of six unarmed Mark 12 reentry vehicles streaking toward their targets on Kwajalein atoll in the western Pacific. The reentry vehicles were originally mounted on two Minuteman III intercontinental ballistic missiles (ICBM's), launched in an operational test from Vandenberg Air Force Base in California. The reentry pattern is typical of a countersilo attack in which two thermonuclear warheads from different missiles are directed toward each hardened target.

cular error probable (CEP): the radius of a circle centered on the average point of impact within which 50 percent of the RV's would fall. Thus the CEP is a measure of the precision with which a given missile delivers its payload.

With estimates of the CEP, the explosive yield and the reliability of a given weapon it is possible to calculate the probability of the weapon's destroying a target of a given hardness. As an example, the Russian missile designated SS-19 Mod (for modification) 3 is estimated to have a CEP of 250 meters, and each of its six warheads has a yield of approximately 550 kilotons. The SS-18 Mod 4 has a similar accuracy and warheads with a similar yield, but it carries 10 independently targetable warheads. Assuming perfect reliability, warheads of this type would have a 63 percent probability of destroying a Minuteman silo hardened to 2,000 p.s.i. Assuming statistical independence, two such weapons would have an 86 percent probability of destroying the same silo. This calculation would indicate that with 2,000 warheads the U.S.S.R. could destroy nearly 90 percent of the 1,000 U.S. Minuteman missiles in their silos. This is the kind of alarming theoretical result that has been given wide circulation.

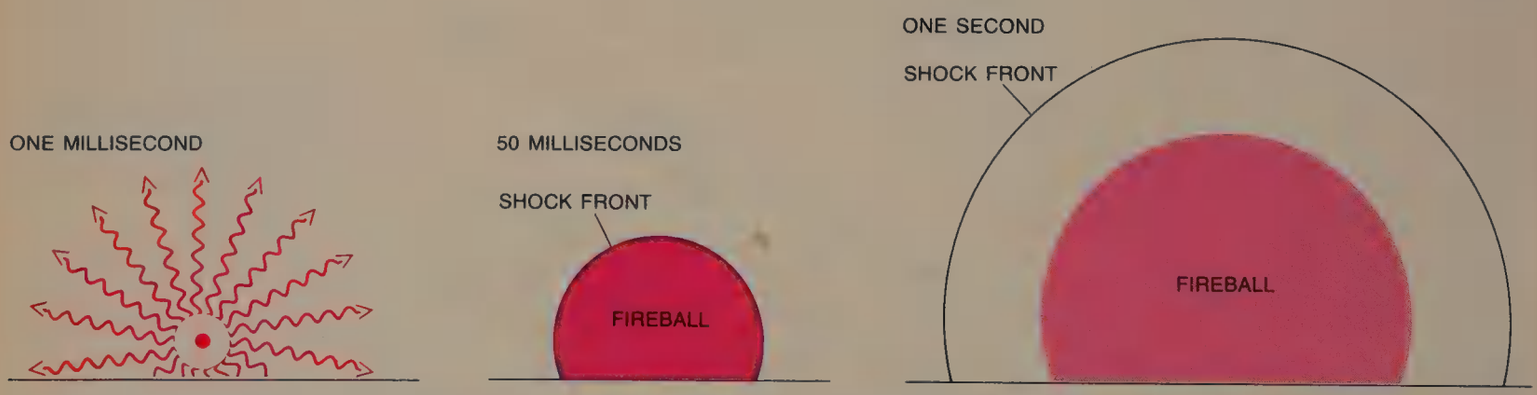
It is unreasonable to expect, however, that a system as complex as a modern ICBM will have a reliability close to 100 percent; a more plausible estimate for the reliability of current Russian ICBM's is 75 percent. Such a reliability would reduce the two-warhead kill probability to 72 percent, which agrees well with the 70-to-75 percent figure currently cited by the U.S. Joint Chiefs of Staff.

Even this figure does not take into account the broad range of uncertainties that must be considered in calculating the possible outcomes of any counter-silo attack. The planner of any such attack will always face further uncertainty about the actual accuracy, reliability and yield of his weapons under operational conditions, and about the hardness of his opponent's silos as well. Several less well known factors, such as interference among the weapons used in the attack, known as "fratricide," will also have significant effects on the outcome. As we shall show, these uncertainties would make it impossible for Russian leaders to have reasonable confidence in destroying significantly more than half of the U.S. land-based missile force, given the current capabilities of Russian ICBM's.

Let us first return to the leading source of uncertainty: the accuracy of the attacking weapons. As we mentioned in passing, the bulk of the errors in an ICBM system will be statistically uncorrelated from one missile to another, contributing to the random distribution described by the CEP. As with most other complex electromechanical systems, however, ICBM's show not only random errors but also systematic ones. Therefore whereas the standard calculations of ICBM vulnerability assume that the center of the impact distribution will be directly on the target, the actual distribution will often be offset somewhat. This distance between the target and the average point of impact is referred to as the bias. For example, if the estimated position of a given target is 20 meters south of its actual position, it is likely that the position of other targets in the same general area, such as other silos in the same missile field, will be similarly misestimated; thus on the average the RV's attacking that field will land 20 meters south of their intended targets. Similarly, the gravitational errors for ICBM's launched from silos close together in the attacker's territory will be strongly correlated, as will the errors due to prevailing winds and atmospheric density variations encountered by RV's reentering the atmosphere over the same area in the defender's territory. On the other hand, it should be possible to remove the most significant biases in the guidance system itself through peacetime testing and analysis.

Any bias smaller than roughly 100 meters has essentially no effect on the success of an attack against silos hardened to 2,000 p.s.i. The reason is that any weapon falling within three times that distance of the target will destroy it. Because of the nature of the bias, however, it will be difficult to place limits on the magnitudes of the bias that might be encountered in an actual attack, owing to the statistical properties of large numbers. If an error is simply random from one warhead to the next, then an attack involving 2,000 warheads would provide 2,000 independent trials for that variable; as with 2,000 rolls of a die, the probability of any significant variation from the average outcome is quite small. On the other hand, if an error is completely correlated over some fraction of the force, then for that fraction the attack provides only a single trial. Just as a roll of two dice might well turn up a two or a 12, rather than the more likely value of seven, so biases significantly larger than 100 meters in a single missile attack cannot realistically be discounted.

There is also some uncertainty about the estimate of the CEP for any given ICBM. A variety of sources of information are available on the accuracy of a given weapon. For example, nondestructive tests on the ground can provide



EARLY EFFECTS of the low-level atmospheric explosion of a 550-kiloton nuclear weapon are represented in the sequence of drawings on these two pages. Both the size of the explosive device and the height of its detonation are assumed to be appropriate for an ICBM attack on U.S. Minuteman silos. In the first millisecond after the warhead is detonated the temperature of the nascent fireball is approximately 400,000 degrees Celsius and the overpressure (the increment above ambient air pressure) is on the order of 100,000 pounds per square inch. Radiation (primarily neutrons and gamma rays) that can destroy a nuclear warhead extends outward to a distance of 800 meters from the point of detonation. After 50 milliseconds the radius of the rapidly expanding fireball has grown to about 500 meters and the temperature inside the fireball has fallen to approximately 75,000 degrees C. The overpressure at the shock front (which at this stage is coincident with the surface of the fireball) is 600 p.s.i. and the wind at the perimeter is blowing outward at several thousand kilometers per hour. After one second the fireball has a radius of 900 meters, an internal temperature of 10,000 degrees C. and a surface temperature of 6,000 degrees. The shock front is now expanding faster than the fireball and has reached a radius of 1,400 meters; at that distance

detailed performance specifications for every component of the guidance system, which can be combined to yield rough estimates of the overall accuracy of the system. The fact remains that many significant sources of error, such as errors attributable to reentry, cannot be realistically tested on the ground. In addition the interaction of the various components of the guidance system in the demanding vibration, shock and acceleration environment of a rocket boost is extremely complex. Hence only a statistically significant number of realistic full-system flight tests can give accurate estimates of the CEP of a ballistic missile.

Such tests are typically divided into two major categories: (1) research-and-



the overpressure is 40 p.s.i. and the wind is roughly 1,200 kilometers per hour. After 10 seconds the fireball has attained its maximum radius of about one kilometer and has begun to rise. The surface temperature of the fireball is approximately 2,000 degrees C. and the radius of the shock front is about five kilometers. (It is visible only at the top of this frame.) The overpressure at the shock front is about 5 p.s.i. Vertical winds at a speed of about 600 kilometers per hour are beginning to suck dust and other debris from the ground into the stem of the ascending cloud. After one minute the characteristic mushroom-shaped cloud has grown to a radius of 2.5 kilometers and its center has reached an altitude of 6.5 kilometers. The fireball has ceased to radiate at visible wavelengths. Vertical winds at several hundred kilometers per hour continue to hold large particles aloft in the cloud and cloud stem. A second-wave reentry vehicle entering the immediate vicinity at any of these early stages stands a good chance of being destroyed by nuclear or thermal radiation, winds or collisions with the larger particles raised by the detonation of a first-wave warhead. Even at later stages smaller particles raised by the first-wave warheads may affect the second-wave warheads (*see illustration on next two pages*). These destructive effects have been termed "fratricide."

development tests, which serve to suggest design changes and to provide initial information about the accuracy the missile can attain, and (2) operational tests, which provide estimates of the accuracy and reliability of the deployed force. Because a single ballistic missile often costs in excess of $10 million, flight-test programs are generally limited by budget constraints. As a result both the U.S. and the U.S.S.R. have tended to perform a comparatively small number of flight tests of each missile system, with intensive engineering analysis of each test. For example, in the initial stages of deployment a U.S. ICBM is typically subjected to between 25 and 30 flight tests, which yield the primary estimate of the system's operational accuracy and reliability. These are followed by from five to 10 operational tests in each year of the system's life cycle, to monitor any changes that might result from prolonged storage.

The operational flight tests of the U.S. are designed to ensure that the tests are as realistic as possible. The ICBM's to be tested are chosen randomly from the operational force. The chosen missile is then brought to alert in its silo, ready for immediate firing; this is intended to provide a test of the crew and the launch-silo electronics. If the missile fails to come to alert, it is listed as a failure and is not tested further. If it passes the test, the missile is removed from its operational silo and is sent to the test range at Vandenberg Air Force Base in California. The RV's are sent to a special facility where the thermonuclear warheads are removed and replaced with telemetry equipment. The missile with the modified RV's is then fired from the test silo at Vandenberg by a crew randomly selected from the operational missile crews, and the RV's reenter the atmosphere over Kwajalein atoll in the Pacific.

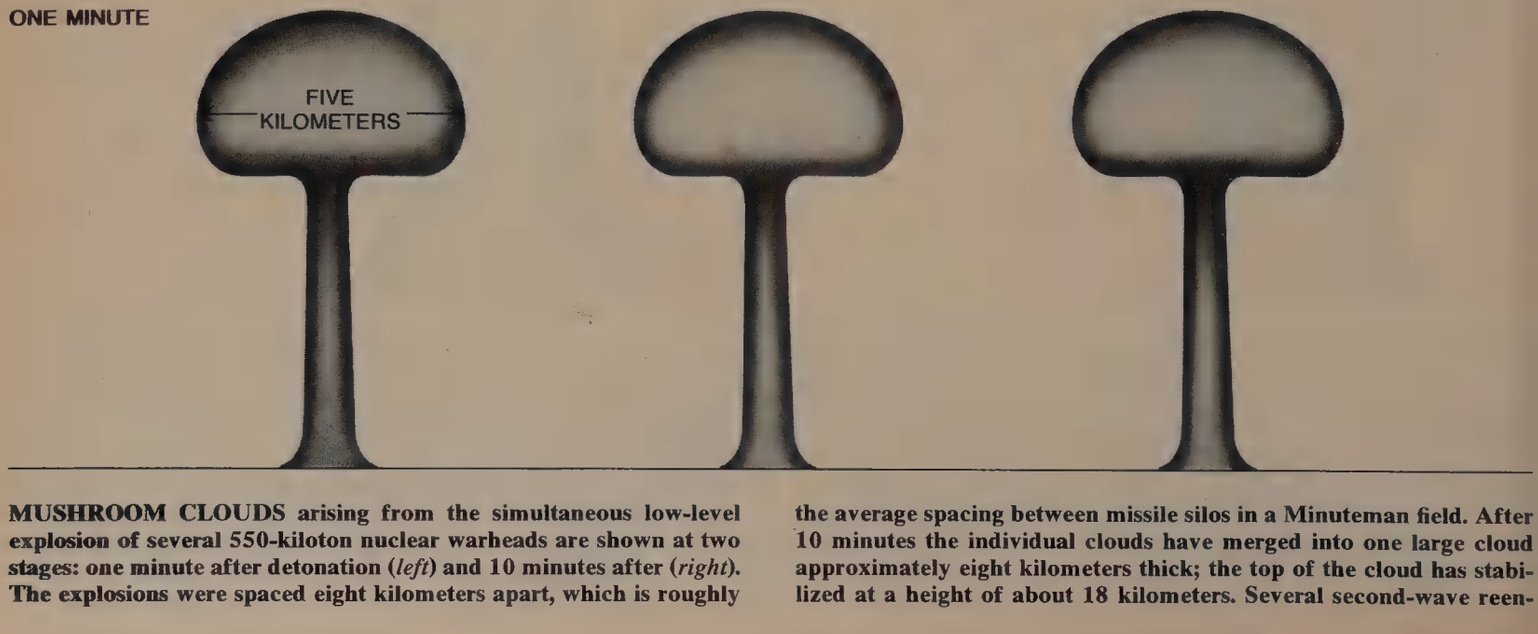
Telemetry equipment on board the missile broadcasts detailed information about such parameters as the flow of fuel, the thrust of the rocket, the performance of the guidance components and the vibration and shocks to which the missile is subjected. This information is picked up by U.S. stations monitoring the tests and by Russian intelligence ships stationed in the Pacific for that purpose. In addition the flight is carefully monitored by radars and optical telescopes based at Vandenberg and on Kwajalein. The information is then intensively analyzed over a period of several months. As a result far more information is generated by a flight test than simply whether or not the missile worked and how far from the intended target it landed.

Significant uncertainties nonetheless persist. First, the number of full-system flight tests is statistically quite small. Second, in spite of efforts to achieve the greatest possible verisimilitude such peacetime testing is still appreciably different from wartime operations. For example, American test vehicles reenter over Kwajalein lagoon, an area where atmospheric conditions are among the placidest in the world. In addition, since the gravity field of the earth varies from place to place, tests over one trajectory cannot in themselves provide estimates of possible gravitational errors over other trajectories.

Perhaps more significant, the range over which most Russian missiles are tested is 6,500 kilometers long, whereas many wartime trajectories would be nearly 10,000 kilometers long. Although the Russians do perform several full-range flight tests in assessing each of their ICBM systems, most of their tests are over the shorter distance. This difference in range has a marked effect on virtually every source of error in the system. Although adjustments to the resulting accuracy estimates can be made, based on a detailed mathematical model of the performance of the system, additional uncertainty will inevitably result. Given these factors, it cannot by any means be ruled out that the CEP in a large-scale countersilo strike could be 10 percent larger (that is, poorer) than the CEP estimated from shots over test ranges; indeed, we believe this is a conservative estimate. For the warheads we have been considering, an unfavorable variation of 10 percent in the CEP alone would reduce the two-shot kill probability against a Minuteman silo from 72 to 66 percent.

Uncertainties in the reliability of an ICBM are separate from uncertainties in its accuracy. Simulations on the ground are considerably more useful for estimating rates of failure than for assessing the accuracy of the weapon system. For example, U.S. Minuteman guidance systems are regularly given simulated flight tests, in which the guidance system is subjected to vibrations and shocks similar to those of a missile flight. As with any sensitive and complex technical system, however, high confidence in estimates of overall reliability can be achieved only through full-system tests, and the number of such tests is quite limited. In addition estimates of the overall operational reliability of ICBM's must take into consideration a broad range of human factors that would be involved in an attack: any large-scale countersilo strike would call for the timely cooperation of several hundred people, whose behavior under such circumstances is unpredictable. Thus 10 percent is probably a conservative estimate of the uncertainty in estimates of overall reliability. Again, an unfavorable variation of 10 percent in the reliability we have assumed for an SS-19 Mod 3 would reduce the number of Minuteman silos destroyed in a hypothetical two-on-one attack from 72 to 67 percent.

Estimates of the explosive power of the thermonuclear warheads have un-



MUSHROOM CLOUDS arising from the simultaneous low-level explosion of several 550-kiloton nuclear warheads are shown at two stages: one minute after detonation (*left*) and 10 minutes after (*right*). The explosions were spaced eight kilometers apart, which is roughly the average spacing between missile silos in a Minuteman field. After 10 minutes the individual clouds have merged into one large cloud approximately eight kilometers thick; the top of the cloud has stabilized at a height of about 18 kilometers. Several second-wave reen-

certainties of their own. This uncertainty is of two interrelated kinds: first, uncertainty about the precise effects of warheads of given yield, and second, uncertainty in the estimates of the yield of a given class of warheads. The peacetime testing of nuclear explosives is limited by considerations of cost, safety, environmental impact, instrumentation and politics. Not least, it is also limited by treaty, including the Limited Test Ban Treaty of 1963 and the still unratified Threshold Test Ban Treaty.

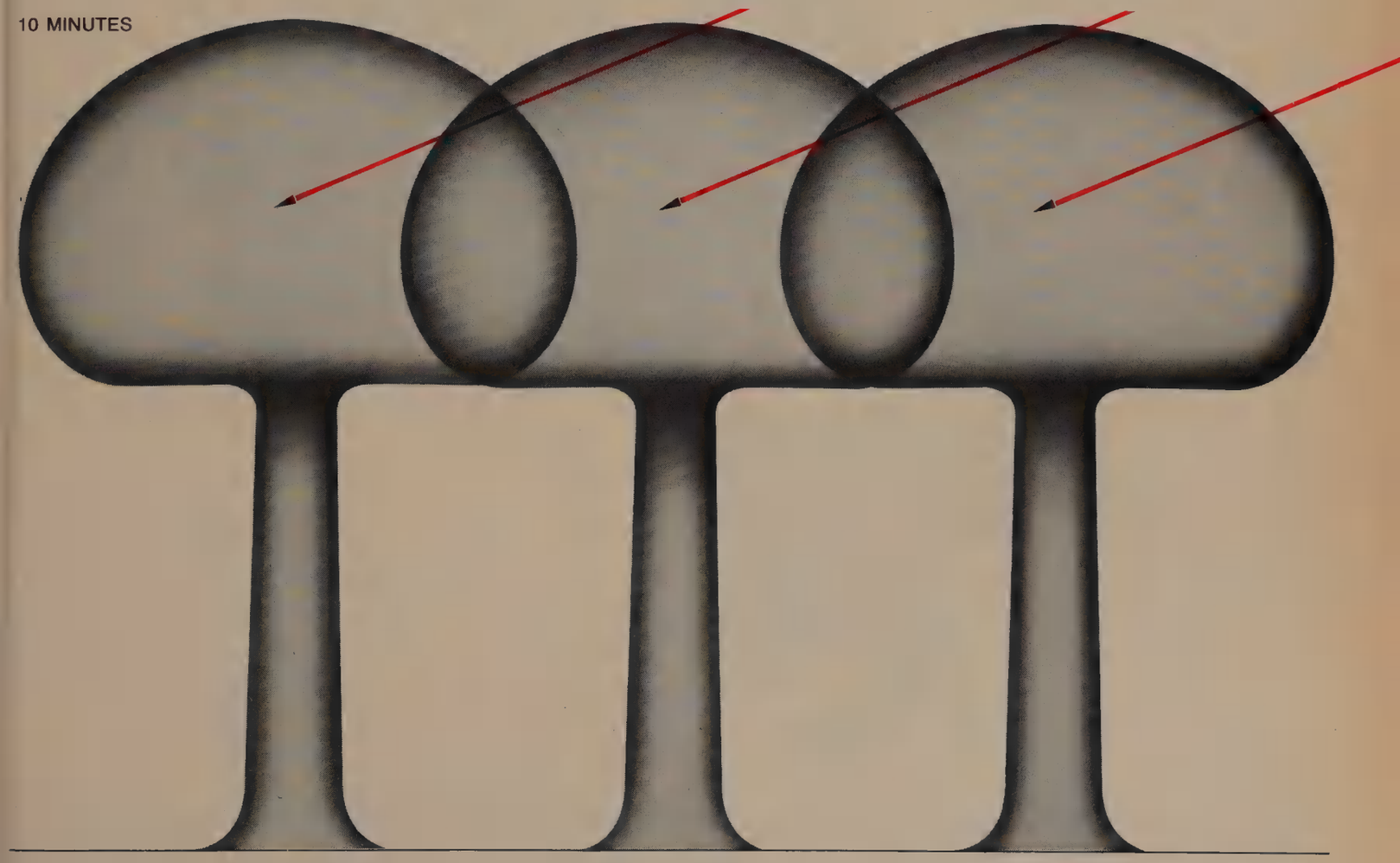
The measurement of overpressures in the extreme range necessary to destroy a modern ICBM silo has been particularly limited, because of instrumental problems and the lack of pressing need. By the early 1960's, when the last American and Russian nuclear tests in the atmosphere were conducted, the hardest targets of interest were roughly an order of magnitude "softer" than current missile silos. As a result there is little in the way of nuclear-test data for overpressures higher than 200 p.s.i., and even less for overpressures higher than 500 p.s.i. The nuclear-test data that do exist for these high overpressures show an enormous degree of scatter, and they often do not agree well with theoretical predictions. Although additional data are available from tests with chemical explosives, large uncertainties are involved in extrapolating such data to predictions of the effects of megaton-range nuclear weapons. The U.S. Defense Intelligence Agency has stated that U.S. estimates of the overpressures to be expected from nuclear explosions in given ranges are uncertain to plus or minus 20 percent.

The second uncertainty in estimating the explosive power of thermonuclear warheads, the uncertainty in estimating the average yield of a given class of warheads, arises from similar sources: the number of tests of any given nuclear weapon is small, and there are substantial instrumental difficulties in measuring the energy yielded by an underground nuclear-test explosion.

The yield to be expected from a given warhead design can be estimated theoretically, but an example of the pitfalls of this approach is provided by the warhead developed for the Mark 12A reentry vehicle recently deployed on the U.S. Minuteman III missiles. The first three tests of the weapon revealed that its yield was considerably less than had been predicted, and the original design had to be modified until, in a fourth test, the weapon achieved its full yield. Although this may be an extreme case, the uncertainty in the average yield of a given class of warheads will be at least 10 percent.

For the sake of simplicity we shall combine the two types of uncertainty in the explosive power of a warhead, describing them both as variations in effective yield. Since the peak overpressure is roughly proportional to the explosive yield, a 20 percent variation in overpressure would result from a 20 percent variation in explosive yield; a conservative estimate of the total uncertainty in the effective yield of a given weapon might then be on the order of 25 percent. An unfavorable variation of 25 percent in the yield of the weapons involved would reduce the effectiveness of the hypothetical attack we have described from 72 to 66 percent.

Although the accuracy, reliability and yield of ICBM's are uncertain, they are at least subject to peacetime testing by a nation that might be considering an attack. This does not hold for the hardness of the silos to be attacked. The overpressure at which a silo will fail depends primarily on the technical characteristics of the reinforced-concrete door at the top of the silo, and it is extremely difficult to obtain reliable intelligence on these characteristics. Indeed, from an attacker's point of view this may well be the largest uncertainty of all. Even U.S.



try vehicles aimed at more-distant targets are shown traversing the cloud blanket. High-speed collisions with comparatively small particles in the clouds could have catastrophic fratricidal effects on the second-wave reentry vehicles, severely degrading their accuracy and perhaps even destroying them. ICBM's that survived the first-wave attack could be safely launched through the cloud cover before the second wave of incoming warheads could safely enter the target area. Hence the attacker might succeed only in destroying empty silos.

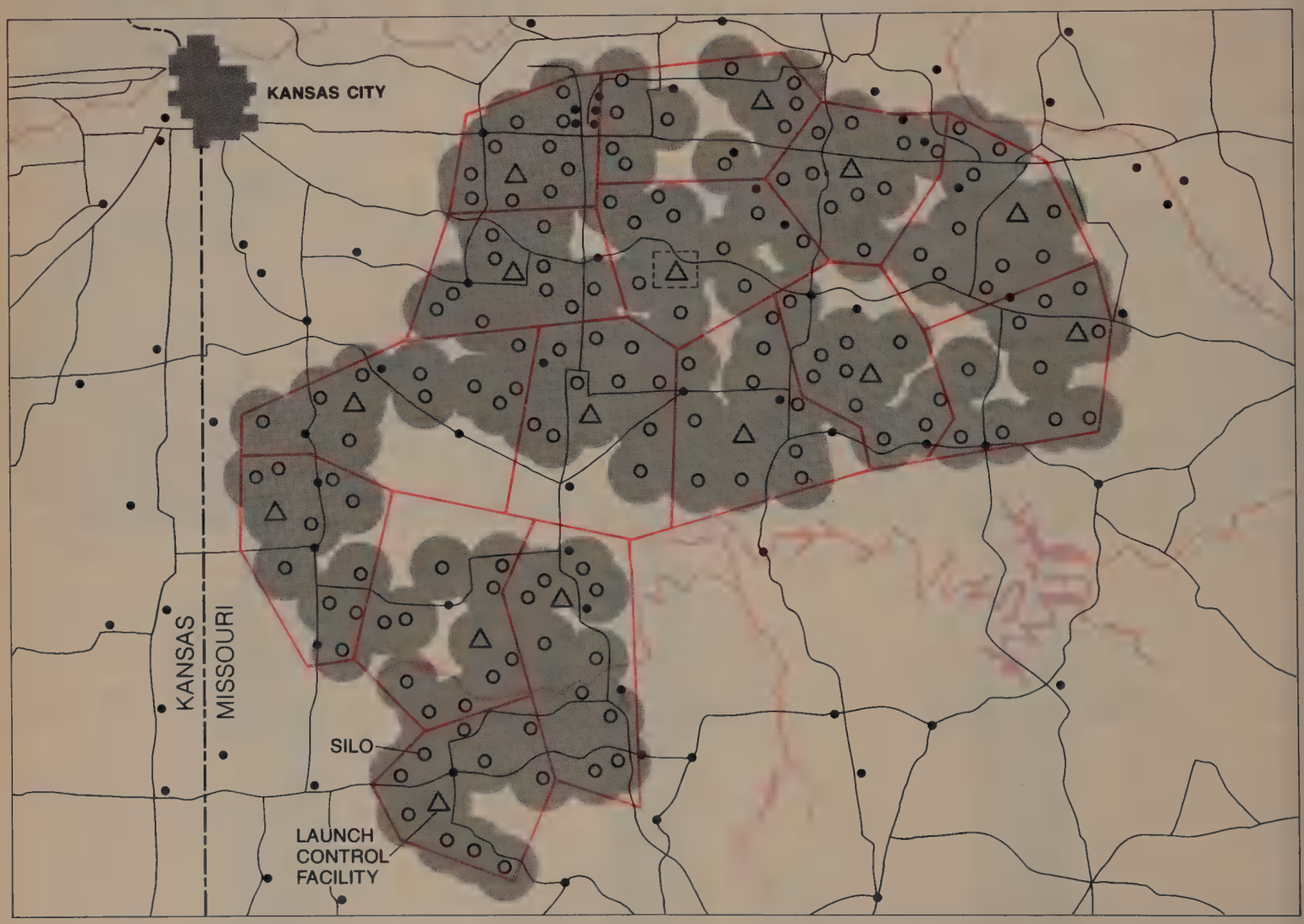
estimates of the hardness of U.S. silos include significant uncertainties; the fact is that no silo has ever been exposed to a nuclear weapon in any test. Estimates of silo hardness are based entirely on theoretical structural considerations and tests of scale models with chemical explosives. Thus from an attacker's point of view the uncertainty in the hardness of the silos to be attacked is likely to be at least 20 percent, if not considerably more. If the silos under attack were capable of withstanding overpressures 20 percent higher than expected, the effectiveness of the hypothetical attack we have been describing would be reduced from 72 to 68 percent.

Another significant uncertainty arises from the interference among the hundreds of warheads that would be involved in an attack. Up to this point we have been assuming that the warheads involved in a countersilo attack would be statistically independent of one another, in other words, that the detonation of one warhead would have no effect on other warheads participating in the attack. This is not at all the case: thermonuclear explosions can have extremely destructive effects on other reentry vehicles, constituting the phenomenon ironically termed fratricide.

In its first milliseconds a thermonuclear explosion gives off an intense burst of radiation, including neutrons, X rays and gamma rays, which in turn generate a powerful electromagnetic pulse. This short-lived burst of radiation is followed by the rapid expansion of a fireball of hot, compressed gases. The fireball expands faster than the speed of sound, reaching a radius of several hundred meters in less than a second. Since the superheated gases in its interior are orders of magnitude less dense than the surrounding air, the fireball begins to rise quite rapidly, much like a hot-air balloon. This creates a vertical wind of several hundred kilometers per hour; indeed, the drag of the wind created by a half-megaton explosion is sufficient to hold aloft a two-ton boulder. In the case of a weapons burst at the optimum height for an attack on a silo hardened to 2,000 p.s.i. the fireball will be in contact with the ground for several seconds; as a result the powerful updraft winds will suck up thousands of tons of dust and other debris, creating the characteristic mushroom cloud.

The cloud rises at high speed; within a minute it reaches an altitude of several kilometers. It then slows down, reaching its maximum height some 10 minutes after the detonation. In the case of a half-megaton weapon the top of the cloud will stabilize at an altitude of some 18 kilometers, with the bottom of the cloud roughly eight kilometers below. By 10 minutes after the detonation the cloud will have covered an area of some 100 kilometers; indeed, the cloud from such an explosion will be so large that in an attack on the ICBM silos in a U.S. Minuteman field the clouds from explosions over individual silos will merge into a single blanket of dust over the entire field.

The significance of the dust and other debris raised by the explosion arises



EXTENT OF CLOUD COVER that would be created by the explosion of the first wave of nuclear warheads in a hypothetical counterforce attack on the Minuteman field in the vicinity of Whiteman Air Force Base in Missouri is indicated on this map. By 10 minutes after the detonation of a 550-kiloton nuclear warhead over each of the 150 Minuteman silos and the 15 launch-control facilities in the target area the merged mushroom clouds would have essentially covered the entire field. If the first-wave warheads were detonated at ground level, the resulting cloud cover would contain millions of tons of dust and larger debris. Even at the greatest height possible for an effective attack on such hardened targets with 550-kiloton warheads, the interfering blanket of particles would still contain hundreds of thousands of tons of material sucked up by the rising fireballs. Population centers, main roads and major rivers are also represented.

from the fact that reentry vehicles traverse the atmosphere at extremely high speed. When an RV reentering the atmosphere in the area of an earlier explosion encounters the cloud, it will be traveling some six kilometers per second. Therefore if the RV collided with a particle weighing several grams, it would probably be destroyed outright; such a collision would take place at several times the speed of a rifle bullet. The smaller particles and dust in the cloud could have catastrophically abrasive effects on an RV: the effect of such high-speed passage through the dust cloud would be equivalent to exposure to an extraordinarily powerful sandblaster. The resulting ablation of the RV's nose cone would severely degrade its accuracy; in extreme cases it could cause the RV to burn up. Moreover, the earlier explosion (or explosions) would have completely changed the density and wind profiles throughout the area in ways that are essentially unpredictable and would have drastic effects on the accuracy of reentering RV's.

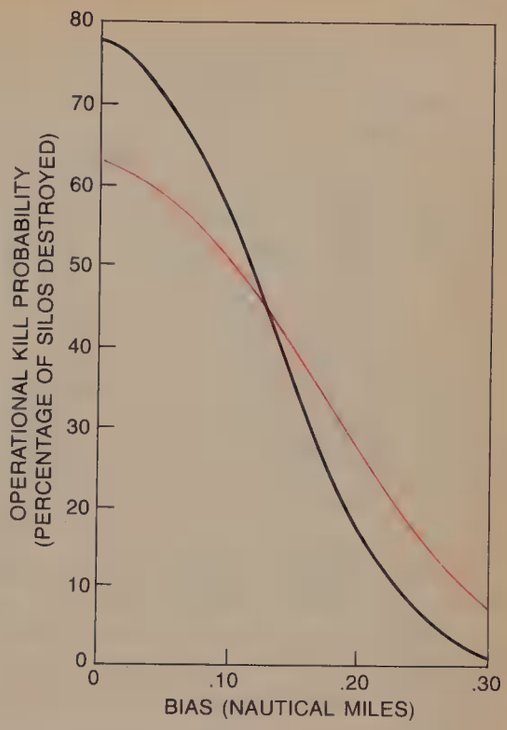
Obviously the problem of fratricide would have to be carefully considered in planning the timing of any attack calling for more than one RV targeted on each silo. If two RV's were targeted on each silo simultaneously, and neither RV failed, then the radiation from the first warhead to explode would destroy the second. If the attack were organized in two waves separated by several seconds, the RV's of the second wave would encounter the fireballs, powerful winds and lethal particles raised by the explosions of the first wave. Although the high temperatures and the outward winds generated by the explosion would subside within the first few seconds, particles large enough to be lethal to an RV would take up to 20 minutes to fall back to the ground, and the clouds of smaller particles and dust would remain until they were dispersed by atmospheric winds. Therefore unless the attacker allowed at least several minutes between the first wave and the second, so that the largest particles would have fallen out of the cloud before the second wave arrived, it seems unlikely that any very large proportion of the second wave would get to its target and do so with the accuracy necessary to destroy it. Even if the attacker allowed several tens of minutes between the two waves, the second wave would still face the dust blanket and the severe atmospheric disturbances created by the explosion or explosions of the first wave.

Since fratricide has never been tested, its effect on incoming RV's cannot be assessed precisely; only the roughest order-of-magnitude estimates are possible. Suppose, in the case of a second wave reentering 10 minutes after the first wave, that passage through the dust cloud and encounters with large particles destroyed only 5 percent of the RV's, and that on the average the effect of the dust and atmospheric disturbances increased the fraction of the CEP attributable to reentry by a factor of two. This, in our view, is a quite conservative estimate of the fratricide such a wave would suffer, given the extreme conditions we have described. These effects alone would reduce the kill probability of a two-wave attack by SS-19 Mod 3 warheads on U.S. Minuteman silos from 72 to 65 percent.

Fratricide introduces an uncertainty that may be even more significant: if the attacker must allow several minutes to elapse between the first and the second wave of his attack, it is quite possible that the ICBM's surviving the first wave will have left their silos before the second wave arrives. Since an ICBM rising out of its silo travels much slower than an RV entering the atmosphere, the particles and dust in the cloud will have comparatively little effect on it. As a result the surviving ICBM's could be safely launched before the second wave of RV's could safely reenter the atmosphere. Although a nearby nuclear explosion that failed to destroy the silo might still keep an ICBM from getting off immediately, in those cases where the first-wave warhead failed the ICBM in its hardened silo would remain undamaged, still available for launching. In the attack scenario we have been considering, involving weapons with a reliability of 75 percent, this would mean that at a minimum 25 percent of the ICBM's could be launched between the first and the second wave of the attack. Combined with the fratricide effects postulated above, this would reduce the fraction of the ICBM silos destroyed in the hypothetical attack we have been considering to 56 percent. Since the vulnerability of an ICBM increases sharply when it leaves its protective silo, however, it is possible that the attacker could keep these weapons from being launched by detonating additional warheads over the silo field at regular intervals so that any ICBM launched from its protective silo would be destroyed; this tactic is called pindown.

Up to this point we have been considering the uncertainties of a countersilo attack on an individual basis. In any real countersilo attack, however, all these uncertainties would be present, making the final outcome of the attack even harder to predict. For example, if all the variables we have been describing were to turn out unfavorably for an attacker, even in the absence of significant bias only 45 percent of the U.S. Minuteman silos would be destroyed. Although it may be unlikely that an attack would be subject to large unfavorable variations in all the uncertain parameters simultaneously, it should be noted that an unfavorable variation in any two of the parameters, combined with fratricide, would reduce the effectiveness of the attack to less than 55 percent, even ignoring the possibilities of bias error and of ICBM's escaping between the two waves. Moreover, we believe the degrees of uncertainty we have postulated are conservative. Thus it would be difficult for a planner to have reasonable confidence that a two-wave attack by SS-19 Mod 3 and SS-18 Mod 4 warheads would destroy significantly more than half of the U.S. Minuteman force.

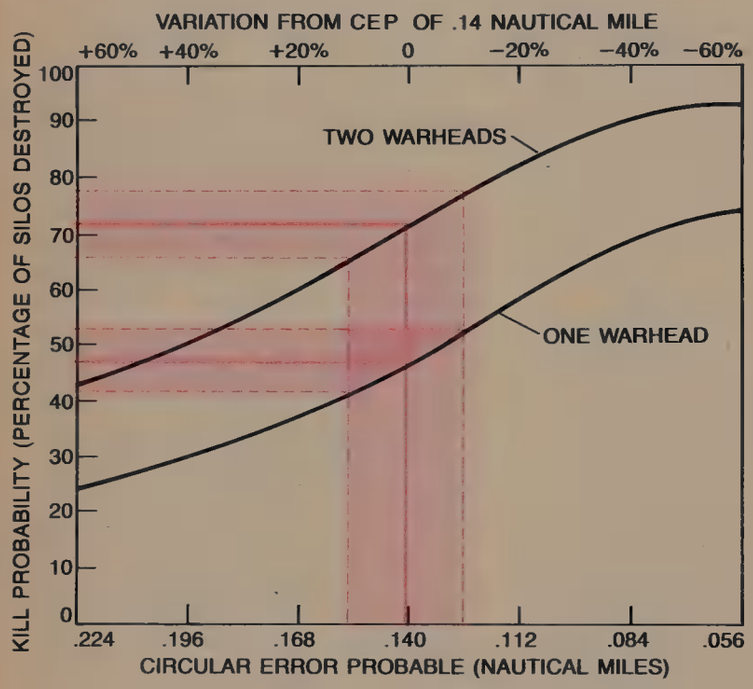
It is possible that a planner would choose to launch more than two warheads against each silo, or to launch larger warheads. In the case of an attack

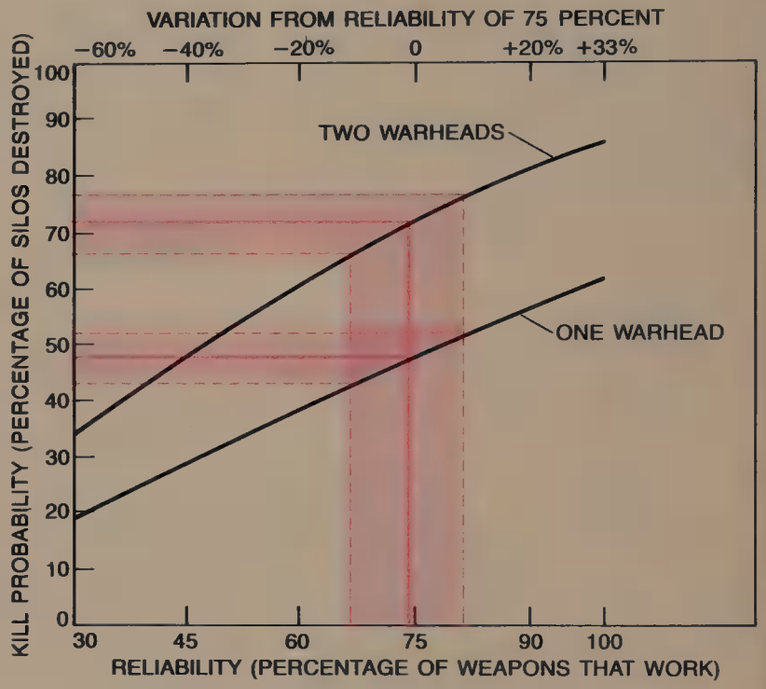


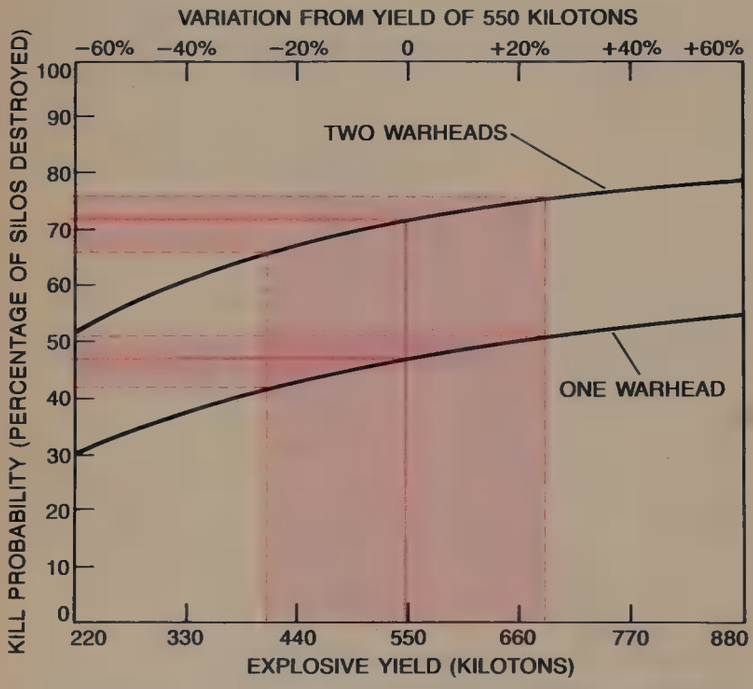
EFFECT OF BIAS on the operational kill probability of an incoming warhead in a hypothetical countersilo attack is shown in this graph for two ICBM reentry vehicles: a Russian SS-19 Mod 3 warhead (*colored curve*) and a U.S. Minuteman III Mark 12A warhead (*black curve*). (Bias is defined here as the distance from the target to the average point of impact of a random sample of warheads aimed at the target.) In both cases the warheads are assumed to have been fired at targets hardened to survive an overpressure of 2,000 p.s.i. In addition it is assumed that each system has a reliability of 100 percent, that the explosive yield of each warhead is 550 kilotons for the SS-19 and 335 kilotons for the Minuteman III and that the circular error probable is .14 nautical mile for the SS-19 and .1 nautical mile for the Minuteman III. (Circular error probable, or CEP, is defined as the radius of the circle within which half of the warheads aimed at the target will fall.) The important thing to note is that for both missiles a bias of less than about .05 nautical mile (roughly 100 meters) would have very little effect on the kill probability, but that a much larger bias could be quite significant.

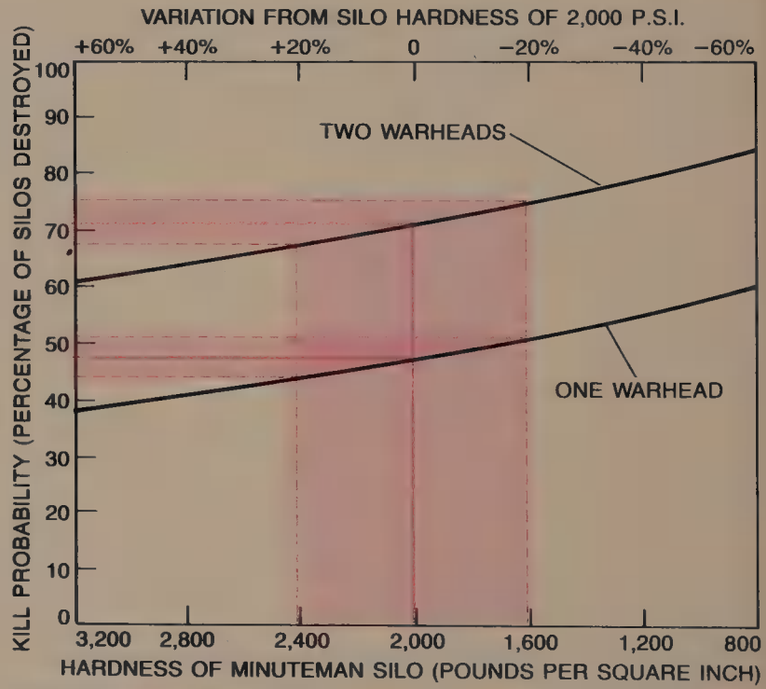
on the U.S. Minuteman force, however, neither of these options would be very attractive. For example, a third wave would encounter fratricide effects from both of the first two waves, and it would allow more time after the second wave during which the surviving ICBM's might be launched. Even if the surviving ICBM's did not escape, a third wave would only raise the overall percentage of the silos destroyed from 45 to 57 percent. Since the 1,000 U.S. Minuteman silos contain 2,100 warheads, this would mean an expenditure of 1,000 additional Russian warheads in order to destroy 250 American warheads; such an attack would disarm the U.S.S.R. faster than the U.S. In addition to its ICBM's armed with MIRV's the U.S.S.R. has 100 SS-19's and SS-18's armed with single large warheads; these, however, are clearly much too few for an attack on the 1,000 U.S. ICBM silos.

For a smaller number of more important targets neither of the two preceding arguments would apply. For example, if the U.S. were to deploy 100 MX ICBM's in Minuteman silos, as has been proposed by the Reagan Administration, the 100 Russian large-warhead missiles could be used to attack them. If these weapons were equipped with the accurate guidance systems now deployed on the MIRVed versions of the same missiles, an attack involving one such warhead on each MX silo, followed by a half-megaton warhead, would destroy 77 percent of the MX force, even allowing for unfavorable variations and fratricide. Since each MX carries 10 warheads, this would mean that even under less than ideal conditions 200 Russian warheads could destroy 770 MX warheads. The expenditure of additional Russian warheads would provide even higher operational kill probabilities.









EFFECT OF INDIVIDUAL VARIATIONS in each of four parameters—CEP, reliability, explosive yield and silo hardness—on the outcome of a hypothetical countersilo attack on the U.S. Minuteman force is projected in this set of graphs. In each case it is assumed that the attack is carried out with one or two waves of SS-19 Mod 3 or SS-18 Mod 4 warheads having a nominal CEP of .14 nautical mile, a nominal reliability of 75 percent and a nominal yield of 550 kilotons. The Minuteman silos are assigned a nominal hardness of 2,000 p.s.i. (Nominal values are indicated by solid-color lines.) The light-color bands show the effects on operational kill probability of a conservatively estimated uncertainty of 10 percent in both CEP and reliability, of 25 percent in explosive yield and of 20 percent in silo hardness.

Thus although the current Russian ICBM force could present a severe threat to a small number of particularly valuable targets, it could not provide a planner with reasonable confidence of destroying significantly more than half of the current U.S. ICBM force. The common practice of citing the probable result of such an attack to two significant figures, with no mention of the attending uncertainty, is grotesquely misleading; all that can realistically be said is that such an attack would *probably* result in the destruction of between 50 and 90 percent of the U.S. ICBM force. We conclude that the magnitude of the threat presented by the current generation of Russian ICBM's has been greatly exaggerated.

This comparatively comforting conclusion may not, however, remain valid indefinitely. Strategic-weapons technology is almost never in stasis, and foreseeable improvements in weapons-delivery systems could drastically alter the situation we have been describing. In the past the accuracy of American and Russian ICBM's has generally improved by roughly a factor of two every seven years; the improvement of strategic-weapons technology may slow somewhat as the room for improvement narrows, but there is little reason to expect that in coming years the pattern will be fundamentally different.

Hence sometime in the late 1980's or in the 1990's the U.S.S.R. may have deployed a force of ICBM's twice as accurate as its current weapons. In addition reentry vehicles specifically designed to penetrate dust clouds are currently under development in the U.S., and possibly in the U.S.S.R. as well. These technological changes could significantly reduce the effect of the uncertainties we have described. If the CEP of the SS-19 Mod 3 were reduced by a factor of two, for example, a two-on-one attack would result in the destruction of more than 80 percent of the U.S. Minuteman force, even allowing for a certain amount of fratricide and unfavorable variations in the attack parameters. In essence, an uncertainty of 10 or 20 percent no longer has a significant effect if the weapon is twice as accurate as it needs to be to destroy its target.

The purely technical uncertainties we have been describing, however, are only the tip of the iceberg. The planner of a real countersilo attack would have to consider a host of other uncertainties, mainly centering on the reaction of the nation under attack. For example, a nation with ICBM's has the option of adopting the policy of "launch on warning" or "launch under attack": launching missiles immediately on receiving intelligence that another nation has launched missiles against it. The U.S. has renounced such a policy, because of the possibility of catastrophic error, but it would be difficult for a planner to have any confidence that this policy would hold in the event of an attack. Nearly 30 minutes would pass between the first detection of a massive launch and the detonation of warheads over U.S. silos; the actions of U.S. policy makers in this period and afterward would be impossible to predict.

If in addition to the U.S. land-based ICBM's the rest of the U.S. strategic nuclear forces are taken into account, a serious problem would arise for the putative planner of an attack. First, the thousands of U.S. nuclear weapons in submarines at sea will be invulnerable to attack for as far into the future as it is possible to predict. Second, it would not be possible to destroy both the U.S. ICBM force and the U.S. strategic bomber force. Part of the bomber force is on a constant 15-minute alert, meaning that the only missiles that could destroy them before they got off the ground would be submarine-launched ones with short flight times. Current Russian SLBM's, however, are much too inaccurate to be effective against hardened silos, meaning that the silos would have to be attacked by ICBM's with longer flight times. If this were done, nuclear weapons would begin detonating over U.S. bomber fields more than 15 minutes before the RV's targeted on the ICBM silos could reach their targets, making it very likely that the ICBM's would be launched before the attacking RV's arrived.

Even if these difficulties could be surmounted, an attack involving more than 2,000 near-ground explosions of megaton-range nuclear weapons would cause between 20 and 40 million civilian casualties. Such an attack cannot realistically be described as "a surgical nuclear strike." It is impossible to believe the U.S. would not respond, relying on some fraction of the thousands of bomber and submarine-based warheads that would surely have survived. In all probability the conflict would quickly escalate to an all-out strategic exchange, which would destroy the attacker as completely as it would destroy the victim of the initial attack.

In the more distant future even these inherent uncertainties might be somewhat reduced by advances in technology. The development of accurate SLBM's, already under way in the U.S., will erode the relation between the alert bomber force and the ICBM force. If extremely accurate maneuvering reentry vehicles are deployed, hardened silos could be destroyed with weapons of much smaller yields, sharply reducing the number of civilian casualties such an attack would inflict. This might at first glance seem desirable, but the possibility of a genuinely "surgical" strike, combined with increased confidence in predictions of its outcome, would increase the temptation to launch such a strike, lowering the nuclear threshold and increasing the probability of nuclear war.

Thus although the current situation is stabler than is commonly believed, the progress of weapons technology bodes ill for the future. It is still possible, however, that stringent limitations on the testing and deployment of ballistic missiles could forestall many of these undesirable technological advances. Far from locking the U.S. into a position of vulnerability, such limitations, if they were effective, could prevent the rapid erosion in U.S. security that will otherwise occur in the years to come. Limitations on the testing and deployment of ballistic missiles could be an important component of arms-control efforts, and they deserve more careful study than they have had to date.

ASSUMPTIONS	BIAS (NAUTICAL MILES)			
	0	.05	.10	.15
100 PERCENT RELIABILITY	86%	84%	76%	63%
75 PERCENT RELIABILITY	72%	70%	62%	50%
LIGHT FRATRICIDE	65%	62%	56%	45%
UNFAVORABLE VARIATIONS	45%	43%	38%	31%

EFFECT OF COMBINED UNCERTAINTIES on the outcome of a countersilo attack is presented in this table. The columns correspond to four different projections of bias. The rows indicate the effects of different assumptions about the technical capabilities of the weapons employed in the attack (and, in one instance, of the hardness of the silos under attack). The first row shows the outcome in terms of operational kill probability of an ideal attack with perfectly reliable weapons on silos with no variation in hardness from the nominal value. The second row shows the results of a somewhat more realistic attack employing weapons with a nominal reliability of 75 percent. The third row shows the effect of a minimal amount of fratricide. The fourth row shows the outcome of an attack in which light fratricide is combined with unfavorable variations (for the attacker) in all the other parameters. In this case the accuracy and reliability turn out to be 10 percent less than the nominal values, the yield of the attacking weapons is 10 percent less and the silos under attack are 25 percent harder than predicted. Even in the absence of significant bias only 45 percent of the Minuteman silos would be destroyed.

Slurry Pipelines

They currently serve to transport metal ores and coal suspended in water. Their greatest potential is for transporting coal from Western states to power stations in other parts of the country

by Edward J. Wasp

Most of the coal in the U.S. (an estimated 72 percent of the minable reserves) is in seven Western states. Although the country is not currently suffering from a shortage of fuels other than coal, experience has shown that the fuel situation can change rapidly, and in any case a cheaper fuel, other things being equal, is always preferable to a costlier one. The reason Western coal is not cheaper than it is currently is the cost of transporting it: most of the electric-power generating stations, which would be the consumers of the coal, are outside those seven states. To build roads and railroads to the mines would call for an enormous capital investment. An alternative that would call for a smaller investment, although still a large one, is slurry transportation: pipelines through which coal is pumped suspended in water.

The longest coal-slurry pipeline currently in operation is the Black Mesa line, which transports coal 273 miles from a mine in northeastern Arizona to a power plant in the southern tip of Nevada. At the mine end the coal is crushed, formed into a slurry and set in motion by three giant reciprocating pumps. At the power-plant end the slurry enters special agitated storage tanks and is fed into continuously operating centrifuges that remove water from it. Out of this process comes a material that looks like damp black beach sand. It is coal for burning in the power plant.

The Black Mesa line has been in steady operation since 1970. It is regarded as the prototype for the much larger systems that will probably be built in the West before the end of the decade. One such system is the one planned by Energy Transportation Systems, Inc. (ETSI), which will move coal from mines in Wyoming south as far as 1,400 miles to Oklahoma, Texas and Louisiana.

Slurry transport is a phenomenon that has been building river deltas and scouring canyons throughout geologic history. It works on the principle that a fluid sufficiently turbulent and fast-moving can transport a particulate solid an indefinite distance. One of the great natural slurry transportation systems is the Mississippi River in the spring: soil is fed into the muddy brown flood from headwaters and tributaries and is carried to the Mississippi delta more than 1,000 miles away. There in an average year the delta is extended six feet by 495 million tons of silt.

These giant natural riverine slurries and manmade pipeline slurries have several things in common. Both consist of finely divided solids suspended in a liquid. In the natural system the solids are ground by erosion; in the manmade one they are ground by machine. Both must be kept moving at or above a critical velocity or they will settle out of suspension. Both eventually do separate out of the transporting fluid. In the natural system they form sandbars or deltas; in the manmade one they are removed mechanically or by evaporating the transport water. Both end up transporting solids from one place to another.

The physics of slurry transport encompasses the principles of fluid dynamics and hydrostatics, whose study goes back at least to Archimedes. For slurry transport, however, the most important scientific findings were those of the 19th-century British physicist Osborne Reynolds. His 1883 paper "An Experimental Investigation of the Circumstances Which Determine whether the Motion of Water in Parallel Channels Shall Be Direct (Laminar) or Sinuous (Turbulent), and the Law of Resistance of Parallel Channels" cleared the way to the understanding, which came later, that turbulence plays a fundamental role in the transport of solids by water. By observation Reynolds developed a formula that generates an index, now called the Reynolds number, indicating whether or not the flow in a pipe is turbulent.

A simple interpretation of the physical significance of the Reynolds number is that the numerator represents the forces tending to cause disorder (higher velocity increases momentum and the degree of randomness in the flow) and the denominator represents the force opposing disorder (greater viscosity damps randomness in the flow). The higher the index, the greater the turbulence, and when in a round pipe the ratio reaches 2,300 (that is, 2,300/1), the flow becomes turbulent. The turbulence in turn increases the capacity of the pipe to carry solids. This last point is easily recognized in the flow of a river: at its head high in the mountains the river flows fastest and carries the coarsest solid material (including rocks); at its mouth it flows slowest and carries only the finest silt.

Modern commercial slurry systems were not developed until after World War II. Since then several commodities have been successfully transported by slurry systems. Iron-ore concentrate, copper-ore concentrate, phosphate-rock concentrate, limestone and the asphalt-like mineral gilsonite are transported by pipeline at several locations around the world. Moscow uses a slurry pipeline to transport garbage from the city to a disposal plant. Moreover, a number of liquids other than water have been examined for slurry-transport purposes: crude oil, methanol (methyl alcohol) and liquid carbon dioxide.

Coal suspended in water, however, was the first slurry system and remains the one with the greatest potential. This is in spite of the fact that it presents special problems. Whereas the metallic ores that have been transported by pipeline are slurried as part of the refining process, coal must be first slurried and then dewatered, steps not normally part of the coal-handling process. Dewatering brings the coal as much as possible back to its original (dry) state for burning in the furnace of a power plant.

The first major experimental work on a coal-slurry system with commercial potential was done in New York City by Wallace C. Andrews, president of the New York Steam Company, which began distributing steam to city residents in 1882. Andrews developed his concept to move coal cheaply and efficiently

from mines to his boiler plant. In the early 1890's he received one patent for his coal-slurry pipeline concept, a second for "An Apparatus for Preparing Coal for Transportation" and a third for "Settling or Storage Pond or Basin for Pulverized Coal." He also built a model coal-slurry pipeline system at his boiler plant at 58th Street and Madison Avenue, complete with pipe running over scaffolding and around corners to simulate cross-country pumping conditions.

It was not until 1914, however, that the first coal-slurry pipeline went into operation. It was in London. Designed by Gilbert G. Bell, it had a pipe eight inches in diameter and a mere 1,980 feet long. It nonetheless did yeoman work, moving 50 tons of coal per hour from barges at a dock on the Thames to a nearby powerhouse.

A major advance then came not from coal and electric-power interests but from the petroleum industry. As the industry expanded, well-drilling technology quickly evolved. The drilling called for methods of preparing and handling drilling muds: slurries pumped down the drill pipe to lubricate and cool the drill bit and to remove rock chips from the well hole. Powerful pumps are needed to force the drilling mud down the drill pipe at pressures approaching 5,000 pounds per square inch. The need for equipment that will operate reliably at such pressures and under the severest conditions has contributed greatly to the development of the massive positive-displacement reciprocating pumps that power slurry-pipeline systems today.

The first modern study of coal-pipeline transportation technology was undertaken in 1951 by the Consolidation Coal Company in Ohio. The company supplied coal from its Cadiz mine in the southern part of the state to the Cleveland Electric Illuminating Company's Eastlake Generating Station 108 miles to the north. At that time coal was shipped a few cars at a time coupled into ordinary freight trains, and the coal company thought it might do better with a slurry line.

Consolidation Coal had to start from scratch, as there were no operating coal pipelines from which to draw technical experience. The coal-slurry research team, of which I was the leader, first built a demonstration plant at Cadiz consisting of two loops of 12-inch pipe. One loop was 4,000 feet long and laid on level ground; the other was 1,200 feet long and laid on an incline. The plant included its own coal-preparation plant, capable of crushing up to 20 tons per hour.

The purpose of the test loop was to examine the distribution and behavior of slurry particles inside the line and to determine the optimum pipeline velocity. If the velocity was too low, the slurry particles would start settling out of suspension, scrape along the bottom of the pipe and erode it. If the velocity was too high, energy would be wasted propelling the slurry and again there would be excessive erosion. The research team also needed to find out if the slurry particle size would be reduced as the slurry traveled down the line. Perhaps most important, the degree to which pressure would drop along the pipeline had to be determined, so that when a full-scale line was built, pump stations would be properly placed and would include sufficiently powerful pumps.

An additional major concern was determining the normal rate of internal corrosion and erosion inside the slurry pipe. The research team needed to find conditions that would allow the pipeline to last for the life of the overall slurry system as well as transport a stable slurry a long distance. To collect the corrosion and erosion data the team designed a short section in the 4,000-foot test loop that included 10 horizontal sections resembling barrel staves. These sections could be removed, weighed and examined after various slurries were run through the loop.

Tests were begun with a slurry in which the coal was in small lumps averaging three-eighths of an inch in diameter. The loss of metal along the bottom of the pipe was enormous compared with that at the top. The plain carbon-steel pipe of the test section was replaced with a stainless-steel one; although stainless steel is slightly less re-



IRON-ORE-SLURRY PIPELINE crosses a gorge of the Savage River in northwestern Tasmania. It carries ore concentrate from the Savage River mine 53 miles over rugged terrain to processing facilities at Port Latta on the coast. Diameter of the pipe is $9\frac{5}{8}$ inches.

sistant to abrasion than carbon steel, it is much more resistant to chemical corrosion. The metal losses were still high at the bottom of the pipe but were not at the top. The high metal losses at the bottom could now be definitely identified as being caused by coal sliding along it. As increasingly fine coal sizes were tried, the coal began to stop dragging along the bottom, and erosion measurements in the test section showed progressively less wear. Furthermore, probes inserted into the pipe near the top, near the bottom and in the middle showed that as smaller particles were slurried the slurry became increasingly homogeneous.

A basic conviction arose that the key to the design of slurry systems that would operate reliably lay not in the selection of exotic pipe materials or the design of special equipment but in the understanding and control of the slurry environment. More specifically, it was thought that if the flow was homogeneous, that is, if the solids were uniformly suspended along the vertical axis of the pipe, the pipeline would be stable with a pressure drop that would be constant with the distance traveled. Two other results would follow naturally from a properly controlled slurry environment. First, if the corrosive environment was controlled and made benign, the wear in the pipe would be uniform and of such a minimal magnitude that a pipe life of more than half a century could be expected. Second, in a homogeneous environment the coal itself would not undergo any attrition, that is, degradation in particle size, during its travel, even over very long distances.

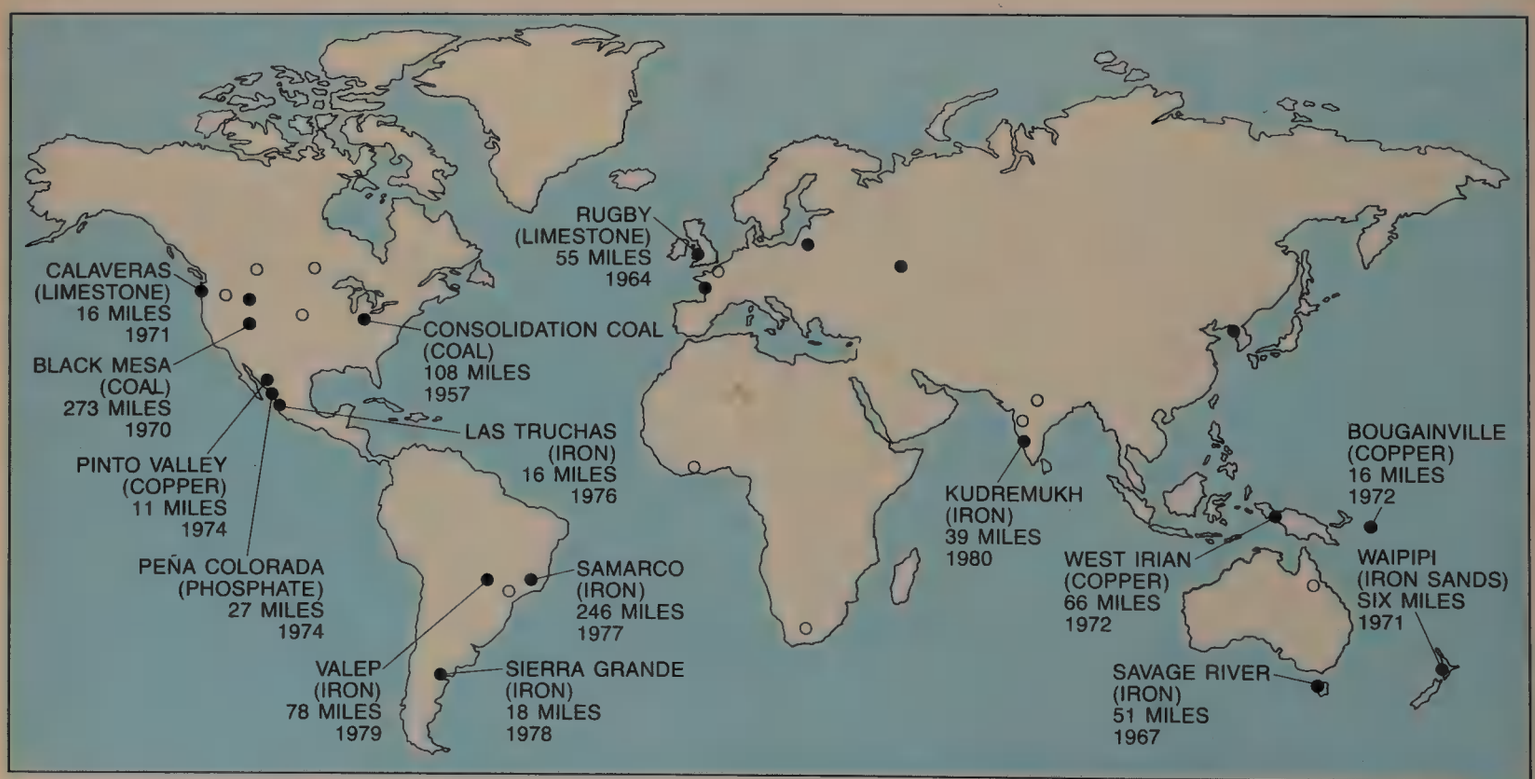
If this basic conviction were proved correct, conventional materials and well-established construction methods that had been developed for pipelines by the oil and gas industry over the first half of the 20th century could serve for slurry-pipeline systems. That could well lead to a rapid, inexpensive startup for a viable coal-pipeline industry.

Several problems remained. Coal particles in the test loop continued to degrade into progressively smaller fines, and the pressure drop in the loop decreased with the time the coal particles spent in it. The research team had to determine whether the level of fines in the full-scale pipeline would increase because of attrition and whether this would affect the pressure drop in the line. Until Consolidation Coal's Eastlake line was completed in 1957 a long-distance, once-through slurry system was not available for experimentation; it was therefore necessary to extrapolate the pressure-drop information obtained from the 4,000-foot test loop to a 108-mile continuous system. Furthermore, since the increase in the level of fines and the decrease in the pressure drop were caused in part by attrition, it was not possible before building the full-scale line to separate the relative contributions to the attrition made by the pipe and by the pump in the system.

To play it safe, the Eastlake coal-slurry pipeline system was designed to accommodate the maximum pressure drop recorded on the test loop and to deal with the fine particles that resulted from the attrition of the coal in the test loop. When the Eastlake system began operating, there was absolutely no measurable attrition in particle size over the 108-mile distance traveled. All the attrition observed in the test loop was caused by a centrifugal pump through which the slurry had to pass more than 100 times as the loop simulated the full-scale system. The pipeline was installed with only three pump stations equipped not with centrifugal pumps but with reciprocating ones.

There was heated debate about what would happen to the turbulence in the pipeline as the slurry traveled down the line away from one of the pump stations. Would it diminish, as some contended, so that the coal particles would lose their support and settle out? Since there was no way to test for the problem until the line was built, the slurry research team had to rely on a "reverse experiment." I reasoned that if the pipeline was shut down and started up again, and if the turbulent forces in the line were strong enough to redistribute the solids uniformly in it before the turbulence generated by the pump reached the sampling point, then the forces in a continuously operating line would be strong enough to maintain the distribution of the solids.

To test this hypothesis the 4,000-foot loop was shut down and then started up. Probes were inserted in the loop at frequent intervals at the top, middle and bottom of the pipe. When the flow was restarted, the course of the slurry resuspension was observed with hundreds of measurements of particle concentration taken at the top, middle and bottom. It was determined that within a minute after the pump started, the slurry had re-



PRINCIPAL PIPELINE PROJECTS are given with their length, the material transported and the date the pipeline went into operation. **The Consolidation Coal pipeline closed down after operating successfully for six years. Circles are where pipelines are being considered.**

suspended all along the length of the loop. After this experiment worry about the effect of distance on turbulence vanished.

The Eastlake pipeline operated successfully for six years until June, 1963, when it was mothballed after having transported some seven million tons of coal. The competition it had given coal-hauling railroads had forced them to operate more efficiently, to develop the concept of the unit train (a train hauling only coal) and to drastically cut their rates for four times the volume of coal that was being moved by the pipeline. In these particular circumstances the railroad had become the cheaper means of transportation.

The next major U.S. coal-slurry pipeline project was sponsored by a railroad: the Southern Pacific Transportation Company. That was the Black Mesa line. Before Black Mesa went into operation, however, several other slurry pipelines were built at widely dispersed locations around the world. The decision to rely on slurry lines at these locations was made easy because material had to be moved across mountainous or otherwise difficult terrain in areas where the construction of roads or railroads would have been prohibitively expensive. Modern pipeline technology made it possible to build the lines on steep grades over routes much more direct than those needed for alternative transportation methods.

At the same time that the Eastlake coal-slurry line was going into operation in 1957, the American Gilsonite Company began operating a gilsonite refinery in Grand Junction, Colo., to process the mineral into coke, gasoline and fuel gas. The gilsonite is mined underground in the Uintah basin of northeastern Utah. High-pressure water jets pulverize the mineral in its seam, and the resulting slurry is pumped to the surface. Once the slurry is aboveground it is moved 72 miles over rugged mountains (including an 8,500-foot pass) through a six-inch pipe to the Grand Junction refinery. If the slurry line had not been built, the gilsonite would have had to be hauled 182 miles from the mine to the refinery by truck or narrow-gauge railroad.



LONGEST SLURRY PIPELINE currently in operation is the Black Mesa line of Southern Pacific Pipelines, Inc. It moves coal 273 miles from the Peabody Coal Company's Kayenta mine in Arizona to the Southern California Edison Company's Mojave power station in Nevada. For most of its length the pipeline is 18 inches in diameter. The Black Mesa line carries 5.5 million tons of coal per year with 99 percent reliability. It went into operation in 1970.

The next major long-distance slurry line was built on behalf of Pickands Mather & Co. in northwestern Tasmania. The company was interested in developing and shipping iron concentrate from the Savage River mine, situated in rugged, semimountainous terrain with few roads and no railroads. The relatively low capital investment needed to construct a slurry pipeline and the low operating costs of this bulk-transportation method convinced the mine's developers that they should rely on a slurry system to ship the ore 53 miles from the mine to processing facilities at Port Latta on the Tasmanian coast.

Major construction barriers had to be surmounted, including two river gorges. The Savage River gorge was more than 1,200 feet wide, and the pipeline crossing was 450 feet above the river. A suspension bridge was built that carried the $9^5/_8$-inch pipe across in a graceful arc. The experience gained in building the Eastlake coal line, including the mathematical formulas and models developed as part of Consolidation Coal's research program, proved to be sound groundwork for the iron-ore line. When the line was first filled with an iron-concentrate slurry, the observed discharge pressure was within 2 percent of that projected by laboratory studies and by mathematical methods based on those applied for the Eastlake line. The Savage River line has been operated with 99 percent reliability for more than 15 years. It moves some 2.5 million tons of iron concentrate per year.

Iron ore is also transported in the Waipipi slurry line in New Zealand. Placed in operation in 1971, the Waipipi Iron Sands Project consists of a relatively short land pipeline and a ship-loading system. A four-mile land slurry line feeds ore to a shore installation. There a 1.8-mile undersea line carries the slurry to an offshore mooring buoy, where special Marconaflo tankers load the slurry and then dewater it on board. After the transport water is removed from the cargo and piped ashore, the tanker sails to its destination with the iron ore. On arrival the ore is slurried again for unloading.

The Savage River iron-slurry line is far exceeded in length by the Samarco iron-slurry line in Brazil, which hauls ore concentrate from mines in the interior through more than 400 miles of jungle terrain to ship-loading facilities on the coast. The 20-inch Samarco line was completed in 1976 and transports 12 million tons of iron-ore concentrate per year.

Two copper-ore-slurry pipelines went into operation in 1972: a six-inch, 17-mile line on the island of Bougainville northeast of Australia and a four-inch, 69-mile line in the West Irian section of New Guinea. The West Irian line receives copper-ore concentrate at an elevation of 8,500 feet and transports it down a mountain to a processing and loading facility on the Arafura Sea.

A limestone slurry is transported by pipeline from the Calaveras limestone mine in California 17 miles through the Sierra Nevada foothills to a cement plant. This major slurry project was completed in 1964 and since then has had an operating availability of more than 98 percent.

Unfriendly terrain and the availability of pipeline technology were important considerations when Southern Pacific Pipelines, Inc., decided to build the Black Mesa pipeline. The line crosses a high desert plateau and four mountain ranges with grades of up to 16 percent along its 273-mile route from the Peabody Coal Company's Kayenta mine to the Mojave power plant of the Southern California Edison Company. Between Pump Station No. 2 at Gray Mountain and No. 3 near Williams, Ariz., the line climbs 3,000 feet in 25 miles. The last 12 miles of the route drops 3,500 feet down

the side of a mountain range into the valley of the Colorado River. For most of its length the Black Mesa pipeline is 18 inches in diameter. On the steep final 12-mile downgrade, however, the pipe's diameter is reduced to 12 inches to diminish the excess pressure caused by the rapid descent.

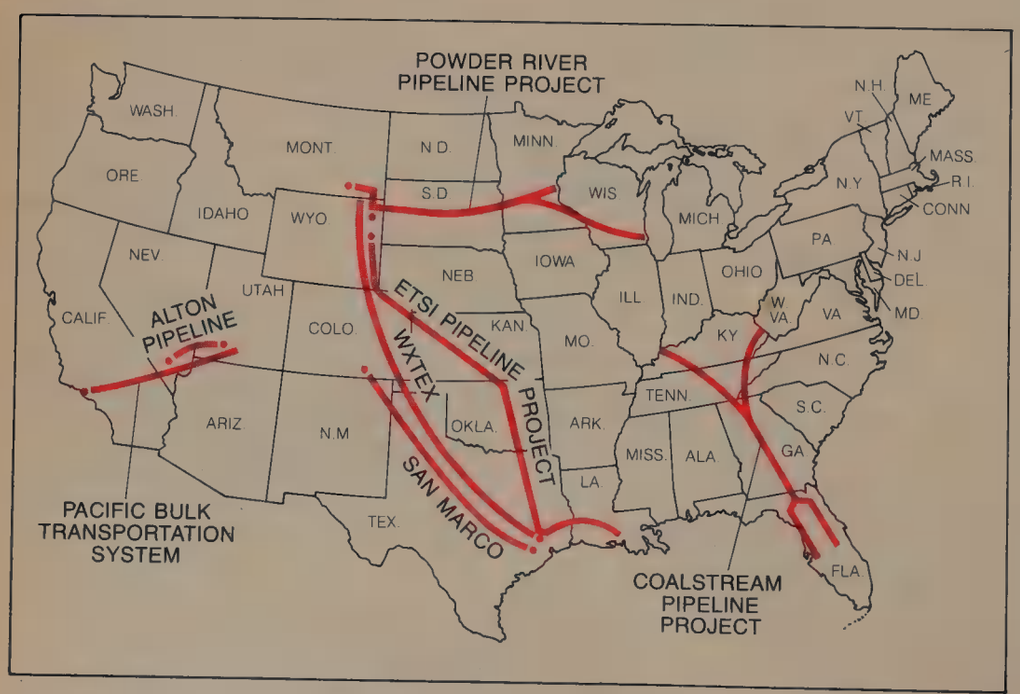
The Black Mesa line is capable of carrying 5.5 million tons of coal per year, and its reliability now approaches 99 percent. The route crosses some of the most scenic country in the American Southwest near the Grand Canyon, Lake Powell and Lake Mead. The local geography and environment strongly militated against the intrusion of a coal-hauling railroad. A pipeline would be out of sight, buried at least three feet underground, making it possible to restore quickly the landscape along the route to its natural condition.

At the Kayenta end of the pipeline the Peabody Coal Company transports two-inch lumps of coal by conveyor belt to the Black Mesa preparation plant. There cage mills crush the dry coal to 3/8-inch lumps. The smaller lumps are then fed to a battery of rod mills that mix the coal with water and wet-grind it until it can pass a No. 14 mesh (a mesh with interstices of 1.19 millimeters). The resulting slurry, which varies slightly in concentration from 46 to 48 percent coal by weight, is first stored in three agitated tanks, each holding a two-hour supply of slurry for the system, that feed Pump Station No. 1.

The positive-displacement pumps at the first station feed the slurry into the main line at 600 pounds per square inch. Pump Station No. 2 at Gray Mountain has four pumps discharging at 1,600 p.s.i. in order to generate the head needed to get the slurry over the Gray Mountain grade. The remaining two stations, No. 3 near Williams and No. 4 near Seligman, are identical with station No. 1. The rates of flow in the pipe are held between five and 5.7 feet per second.

At the Mojave power plant the slurry discharges into one of three agitated tanks capable of holding it for 24 hours. It is then pumped into the centrifuges that dewater it. In addition the Mojave station has eight 75,000-ton Marconaflo ponds designed to store and recover large volumes of slurry coal. There are 20 centrifuges for each 750-megawatt generating unit. The coal cake that is yielded by the centrifuges still contains about 25 percent water. It must be further dried and pulverized before it is blown into the boiler fireboxes with heated air.

At about the same time that the Black Mesa system was being designed and built, an even more ambitious project was envisioned: a massive coal-slurry system linking the newly developed coal mines in Wyoming's Powder River Basin with power stations in states to the south as far east as Texas and Louisiana. Several other coal-slurry pipelines have been proposed, including the Powder River Pipeline, which would transport Wyoming coal to Wisconsin and Illinois; the Coalstream system, which would transport coal from both West Virginia and southern Illinois mines to power stations in Florida; the San Marco system, which would transport Colorado coal to the Houston area; the Alton system, which would transport coal from southern Utah mines to Utah and Nevada power stations, and the Pacific Bulk Commodity Transportation System, which would transport Utah coal to a ship-loading facility on the California coast for export to Japan.



LONGER COAL-SLURRY PIPELINES are scheduled to go into operation by the end of the decade. The longest pipeline is the ETSI Pipeline Project of Energy Transportation Systems, Inc. It will carry coal from mines in Wyoming as far as 1,400 miles into Texas and Louisiana.

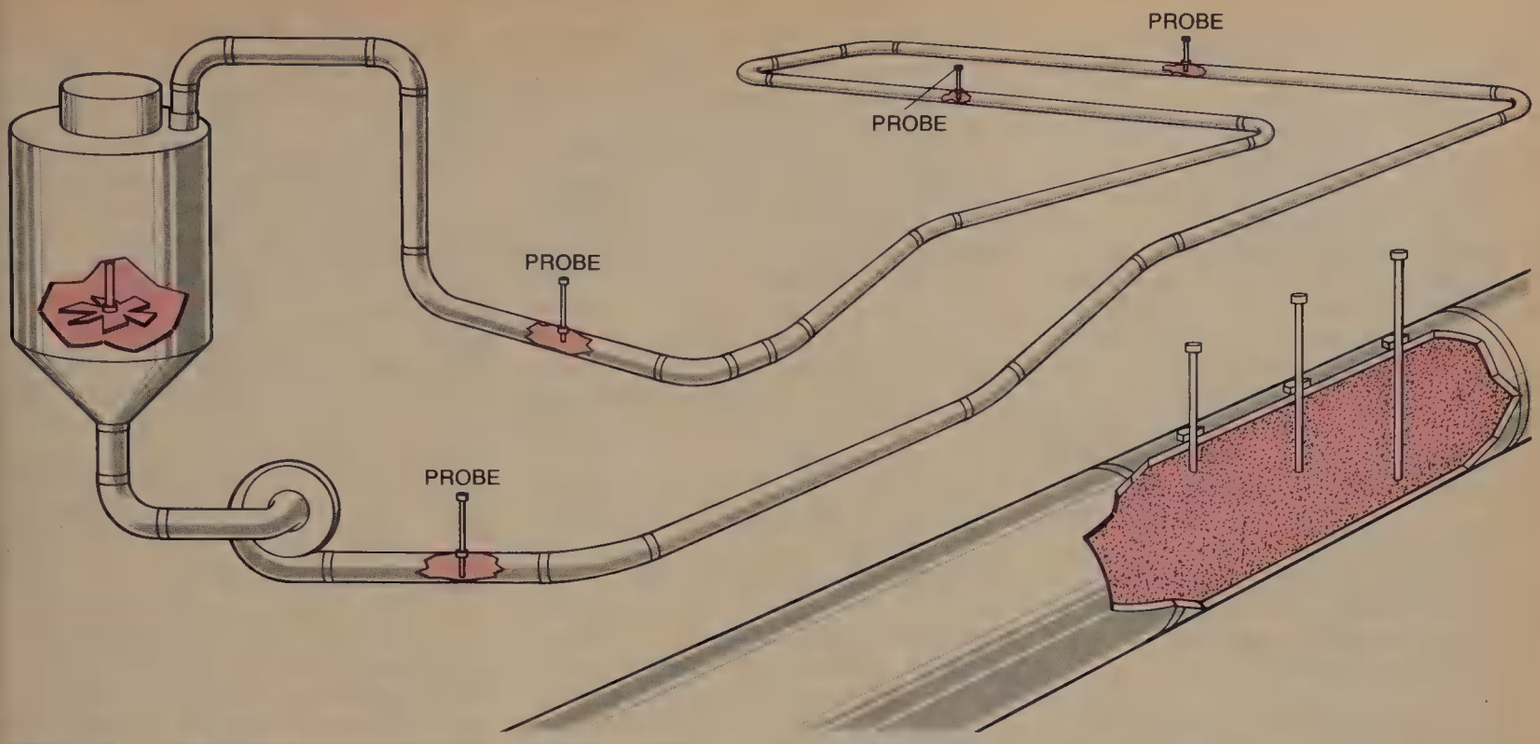
Bechtel Petroleum, Inc., is currently negotiating the construction of two new coal-slurry pipelines in Asia. It is conducting a feasibility study jointly with the China National Coal Development Corporation on a 450-mile pipeline to run from Inner Mongolia to the port of Quinhuangdao and to four power plants. Together with Armand Hammer of the Occidental Petroleum Corporation, Bechtel is also discussing with Russian officials a coal-slurry pipeline from Siberia to Moscow. Proposals for two new coal-slurry pipelines, both about 1,200 miles long, are being studied in India.

To make the pipeline from the Powder River basin to the south a reality Bechtel joined with the investment-banking firm of Lehman Brothers Kuhn Loeb, Inc., to form Energy Transportation Systems, Inc., and began the ETSI Pipeline Project in 1973. Today the joint venture is owned by the Texas Eastern Transmission Corporation, Kansas-Nebraska Natural Gas Co., Inc., InterNorth, Inc., and Bechtel. Lehman Brothers left the joint venture at the beginning of this year.

Of several large-scale slurry-pipeline systems being planned today the ETSI project is the one closest to construction. During its lifetime on the drawing boards the proposed ETSI system has been subject to change, reflecting changing market conditions. As the system was originally conceived, it would include three coal-preparation plants that would receive and slurry coal from several different mines near Gillette, Wyo. Two of the outlying preparation plants would be linked by gathering lines to the third: the Thunder Basin plant at the head end of the main slurry line.

Water would be transported to the preparation plants from the Lake Oahe reservoir on the Missouri River near Pierre, S.D., through a buried 34-inch steel pipeline. The 20,000 acre-feet of water that will be used in the slurry pipeline to transport 30 million tons of coal per year is less than .1 percent of the annual flow of the Missouri at Pierre. The 30 million tons of coal per year represents about 4 percent of all U.S. coal production.

At the coal-preparation plants the same type of cage mills and rod mills installed at Black Mesa will grind and slurry the coal received from the mines by conveyor belt. The slurry will consist of approximately 50 percent coal and 50 percent water by weight. The main line will be a carbon-steel pipe 40 inches in diameter between its Wyoming end and the first delivery points in Oklahoma. Electrically powered 1,500-horsepow-



TEST LOOP was needed before construction could start on the first coal-slurry pipeline: one from the Consolidation Coal Company's mine in Cadiz, Ohio, 108 miles north to a generating station of the Cleveland Electric Illuminating Company. Actually two loops were built, one 1,200 feet long and the other 4,000 feet long. The slurry was kept agitated in a tank. Probes along the loop monitored velocity, pressure and particle distribution. Sections of the pipe could be removed to determine whether the flow caused wear. In this way optimum particle size, particle concentration, velocity and pressure could be established. Full-scale pipeline went into operation in 1957.

er positive-displacement reciprocating pumps in stations from 80 to 100 miles apart will be used to propel the slurry through the line at approximately the walking speed of a man, which is required to keep the slurry in turbulent flow. The slurry will take about two weeks to travel from one end of the line to the other. The line will carry the slurry almost due south across Wyoming, then turn southeast across the northeast corner of Colorado before crossing Kansas in a southeasterly direction into Oklahoma.

When the line reaches Oklahoma, it will split. One leg will head almost due south into Texas. The other will serve power stations in northern and northeastern Oklahoma. Alternative configurations to serve additional power stations in Texas and Louisiana are being studied and may be included in the final system, depending on the outcome of ETSI's marketing efforts.

Coal from specific mines will be transported in batches to specific generating stations to meet the ash, sulfur and B.t.u. requirements of their boilers. Each public utility in the ETSI system would have a contract to buy coal from individual Powder River basin mines. To ensure that the right coal reaches the right destination, batches of coal will have a recorded history prepared when they arrive from the mines at a coal-preparation plant. The history will include an analysis of the batch's chemical properties and heat content. The information will be compared with that for batches arriving at each terminal facility. If the analysis of the arriving batch matches that of the batch expected for delivery, the arriving batch will be delivered to the dewatering plant for processing.

To aid in tracking individual batches through the main slurry line electrolytic markers consisting of a salt solution will be placed in the line ahead of and behind each batch. When the markers pass a pump station, a dewatering terminal or some other way station on the line, they will cause the electric potential across the line (measured by resistance meters) to increase sharply, signaling the beginning or the end of a batch. A group of agitated tanks at each terminal facility will enable the terminal operator to choose between processing the slurry immediately or storing it for later dewatering or later shipment to a downstream terminal in a predetermined sequence.

The entire cycle of coal preparation, pipeline transportation and dewatering has been thoroughly and successfully evaluated with more than 1,000 hours of operation at ETSI's Coal Evaluation Plant in Arkansas. The Coal Evaluation Plant was constructed for full-scale equipment test, so that the behavior of various coals could be studied through the grinding, slurrying, shipping and dewatering processes. The evaluation program demonstrated two things. The first is that coal slurry meeting pipeline specifications can be produced with commercially available cage mills and rod mills, each mill consuming about the same amount of energy as the mills at the Black Mesa coal-preparation plant. The second, at the other end of the circuit, is that the Coal Evaluation Plant consistently produced dewatered slurry coal having between 7 and 11 percent surface water with the reliability needed for commercial operation.

Through the use of a continuous-belt filter press in the dewatering cycle, the dewatering process recovered 100 percent of the coal from the transport water; no more than 30 parts per million of suspended solids remained behind in the water after the coal was removed. The dust-collection equipment on the vibrating-bed dryers worked more efficiently than had been expected. This fact, together with the complete recovery of the solids from the water, assures that both Federal and state environmental standards for industrial air emissions and wastewater discharges at the dewatering plants of the system will easily be met.

Western coal has one disadvantage in addition to its being far away from most existing power stations. That disadvantage, like the one of remoteness, is negated by slurry-pipeline technology. It consists in the fact that Western subbituminous coal has a relatively low heating value and a high inherent moisture content. A hydrothermal coal-treating process to upgrade B.t.u. content and reduce moisture had been explored on a small scale in the laboratory. This treatment was found to permanently raise the B.t.u. content and reduce the inherent moisture content of coal by heating it in slurry form to above 600 degrees Fahrenehit. To find out if this heat treatment could be applied in a large-scale continuous process, Texaco, Inc., joined

with Bechtel in running a 20-ton-per-day pilot plant at Montebello, Calif.

The coal to be processed was first ground and then slurried in water at concentrations of between 45 and 50 percent. The slurry was heated to between 500 and 650 degrees F. in heat exchangers pressurized to keep the slurry liquid. The resulting dewatered coal had a higher heating value than the coal fed into the process because much of the volatile matter had been driven out of it. The B.t.u. content was raised by up to 25 percent; the final test run increased the heat content of a pound of Western coal from 8,500 B.t.u. per pound to 10,500. (High-volatile Kentucky bituminous coal has a heat content of 14,800 B.t.u. per pound.) The process could easily be adapted for minehead coal-preparation plants or dewatering terminals on slurry systems.

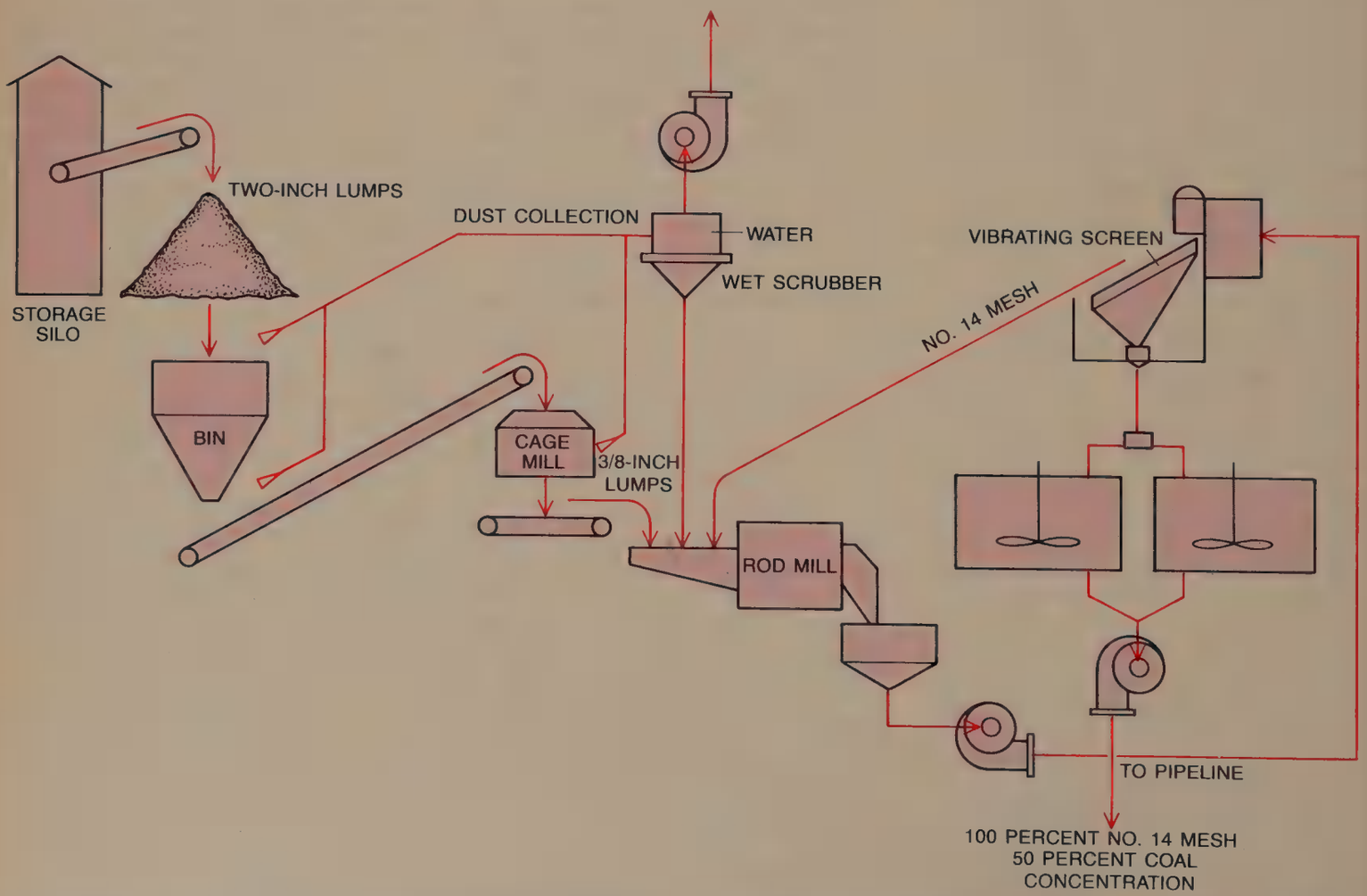
The distance penalty for the use of Western coal is outweighed by a single great advantage: it has the lowest sulfur content of any coal found in the U.S., making it ideal fuel for today's environmentally clean generating stations. Sulfur dioxide is the chief pollutant in coal smoke and may be the cause of "acid rain." Modern coal-fired generating stations rely on scrubbers to remove sulfur dioxide from their stack gases. The less sulfur there is in the coal, the less difficult it is for the scrubbers to clean up the sulfur dioxide.

Two of the major successes of the ETSI project are the experience with the coal-evaluation plant and the results of engineering design. Others include the acquisition of about 90 percent of the rights of way needed for the main slurry system, the acquisition of 70 percent of the easements needed for the water-supply line, the securing of a stable supply of water and the granting of a favorable environmental-impact statement by the Department of the Interior.

In order to acquire the rights of way opposition from railroads had to be overcome. The railroads' major tactic has been to refuse ETSI crossing rights under their tracks. At first this appeared to be an obstacle impossible to overcome. Legal research into the original granting of Western railroad rights of way more than 100 years ago revealed, however, that in many places the railroads had only a surface easement for their tracks; the subsurface rights were retained by the landowners. Title searches made in counties through which the ETSI pipelines would pass located these "windows," and ETSI negotiated easement agreements with the landowners. Then, through a series of court cases filed to validate the agreements, ETSI overcame the railroad-crossing obstacle. Of the 65 window cases filed in the lower courts, all were decided in ETSI's favor. Four subsequent railroad appeals to higher courts were dismissed, and ETSI is now free to build its pipeline under railroad tracks without getting any further approvals of window crossings.

If there had been Federal eminent-domain laws for coal-slurry pipelines, ETSI's window program would have been unnecessary. Such laws are vitally needed for the construction of the other planned coal-slurry systems in this country, particularly those crossing Eastern states where railroads hold full title to their rights of way.

The first coal-pipeline legislation was proposed in 1962 by President Kennedy. His bill, which would have inaugurated the building of a coal-slurry pipeline from West Virginia to New York and New Jersey, was opposed by Eastern railroad companies and was defeated. In 1974, at the height of the Arab oil embargo, a second coal-pipeline bill was introduced in Congress. This time, still opposed by railroad companies, it passed the Senate by voice vote but was defeated in the House Committee on In-



SLURRY IS PREPARED at the mine end of a coal-slurry pipeline. Typically the coal is first crushed into two-inch lumps (*top left*). It is reduced to smaller lumps in a cage mill, water is added and it is further crushed in a rod mill. It is then passed through a mesh to ensure that no large lumps remain in it. The coal slurry is kept suspended in agitated tanks (*bottom right*) until it is pumped into the pipeline.

terior and Insular Affairs. A third bill was introduced in 1975, only to be tabled in committee the following year. The coal-pipeline bill of 1977 reached the House floor and was defeated. Last year legislation cleared House and Senate committees but did not come to a vote before Congress adjourned. In September of this year similar legislation was defeated on the floor of the House. This, however, does not affect the eminent-domain arrangements that will enable ETSI to go forward.

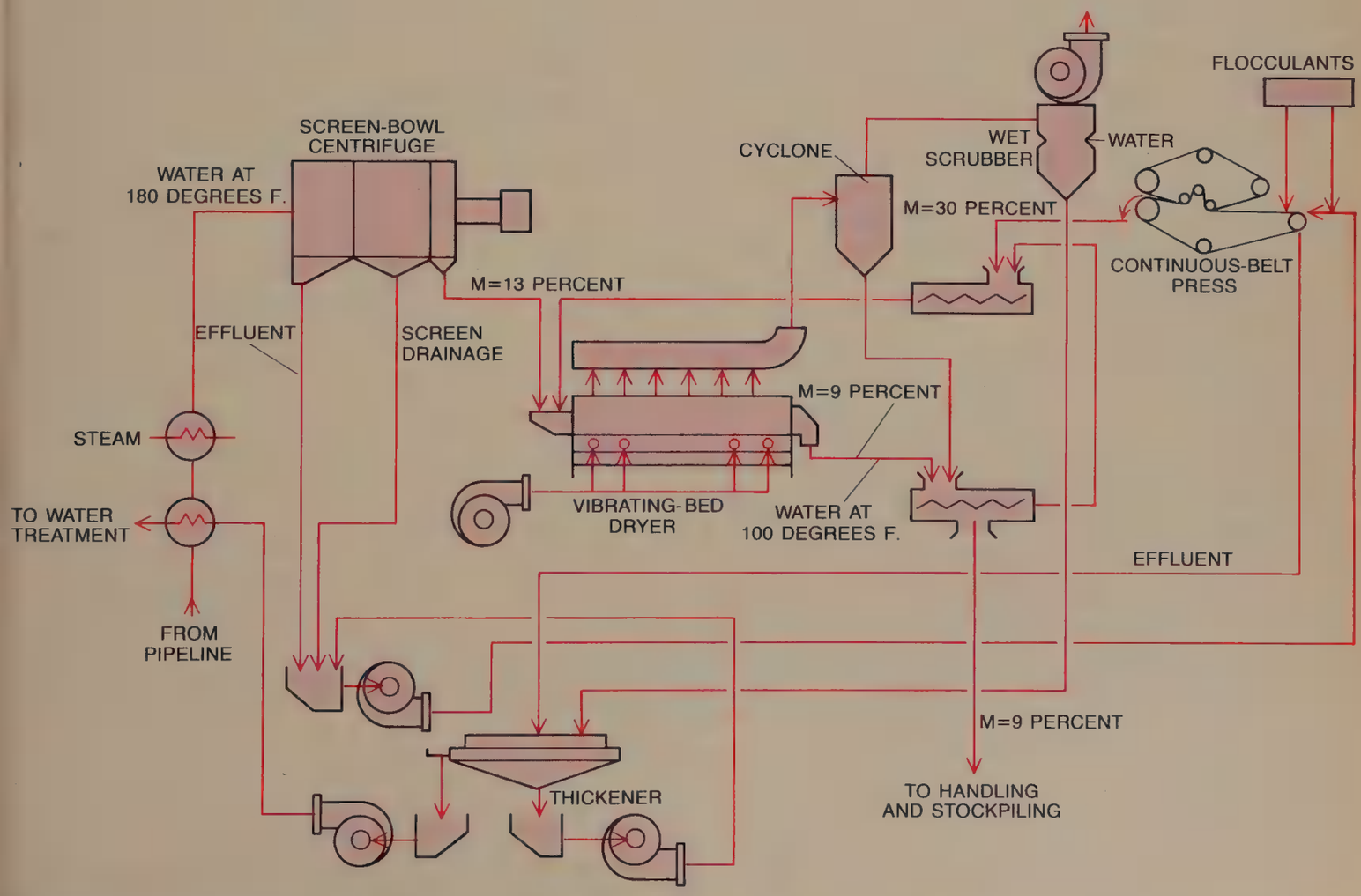
By the end of March of this year ETSI had acquired a total of 1,161 miles of right of way for its main slurry line and 174 miles of right of way for the South Dakota water line, the results of the largest right-of-way acquisition program ever undertaken at one time in the contiguous 48 states. Of the total rights of way acquired only 58.7 miles were across Federal lands and only 40 miles were across state lands.

The Federal lands it is necessary to cross are chiefly the national grasslands in Wyoming and Colorado and other lands in the same states administered by the Federal Bureau of Land Management. It was for this reason that ETSI needed to have an environmental-impact statement. The statement was issued by the Department of the Interior after two years and $6 million in studies. It found the proposed ETSI system was both environmentally acceptable and a "cost-competitive option for transporting coal." It noted that the buried pipeline would have a minimal effect on the environment and that vegetation would restore itself over the right of way within a few years.

In December, 1981, ETSI signed an agreement with the state of South Dakota to buy up to 50,000 acre-feet per year from Lake Oahe. The lake is one of several "main stem" reservoirs built and managed by the U.S. Army Corps of Engineers under the Pick-Sloan plan to end flooding along the Missouri River. At present no use is being made of the water stored behind the Oahe Dam, although Federal and state studies indicate that three million acre-feet per year is available for industrial purposes. As part of its agreement with the state ETSI has agreed to transport 4,300 acre-feet of water per year free of charge to communities and rural water districts along its aqueduct route in western South Dakota.

When construction on the ETSI system begins, the work will proceed simultaneously on the Oahe water line, the main slurry line, the preparation plants and their gathering-line systems, and the dewatering terminals at ETSI customers' generating stations. ETSI expects that as many as 5,500 workers will be employed building the system. The entire effort will create more than 12,000 jobs on the system during construction and some 24,000 additional jobs in related industries. The value of the materials and equipment needed to build the system is expected to total more than $1.5 billion out of a total cost of about $3.5 billion, a cost estimated to be 9.5 percent of the $36.1 billion in commercial and industrial construction contracts awarded in the U.S. in 1982.

Black Mesa showed how to build and run a reliable long-distance slurry pipeline. The ETSI system would dramatically scale up the Black Mesa concept into a pipeline system carrying slurry across a third of the country. One would have to go back more than 100 years to the construction of the transcontinental railroad to find a transportation project equal in size and potential impact on the U.S. economy. There is every reason to believe that because coal-slurry pipelines will contribute to the economy and secure energy supplies for the future the necessary legislation will finally be written into law and the infant industry will be able to grow to adulthood.



SLURRY IS DEWATERED at the power-plant end of the pipeline. The coal slurry coming in from the pipeline is first steam-heated to 180 degrees Fahrenheit. It is run through continuously operating centrifuges that remove all but 13 percent of its surface water (*M*). Other components of the system remove an additional 4 percent of the water. The water removed meets all Federal environmental standards.

Centaurus A: the Nearest Active Galaxy

Active galaxies are galaxies that radiate as much as a million times more energy than typical galaxies. The question of what powers the radiation is studied by observations of Centaurus A

by Jack O. Burns and R. Marcus Price

Among the largest and most intriguing objects that have yet been observed in the universe are the objects known as active galaxies. Active galaxies, which make up only a few percent of all known galaxies, can emit a million times as much energy in the form of electromagnetic radiation as an ordinary galaxy. The emissions from an active galaxy extend over many different frequencies, including gamma rays, X rays, visible radiation, infrared radiation and radio waves. Much of the recent observational work has been done at radio frequencies, where the emissions from an active galaxy can be 100,000 times as strong as those from an ordinary galaxy.

The active galaxy nearest our own galaxy is Centaurus A, which is about 15 million light-years away. The visible outline of Centaurus A is a few tens of thousands of light-years across. Extending at an angle from the central plane of the visible structure, however, is an elongated radio envelope some three million light-years from tip to tip. If the radio-emitting region of Centaurus A were visible to an observer on the earth, it would appear to be 20 times as wide as the moon.

How is the enormous radio structure of an active galaxy such as Centaurus A created and maintained? The combination of observations made at X-ray, radio and optical frequencies is beginning to tell astronomers a great deal about how Centaurus A is organized. The material of the radio structure is expelled like a stream of water from a hose by an engine at the center of the galaxy. From the engine comes a thin beam of radio plasma: a highly ionized gas whose main components are electrons (moving with nearly the speed of light), magnetic fields and possibly protons. As the electrons travel outward from the center of the galaxy they emit radiation with a wide range of frequencies. In doing so they lose energy, and their motion would soon stop if they were not continuously supplied with energy. As the beam moves outward from the engine it passes into three radio lobes arranged in order of increasing size. The middle lobe coincides with a series of optical filaments, thin regions that give off radiation in the visible range. The filaments have recently been shown to be regions where new stars are being born.

Although much has been learned about active galaxies in the past decade, some of their most fundamental features are not understood. The nature of the engine that drives the plasma beam is not known with certainty and the details of the mechanism can only be guessed at. Furthermore, there are several possible explanations of how the electrons in the beam are accelerated as they move outward from the galactic core. We cannot choose the correct explanation on the basis of current knowledge. We also do not know how the radio and optical components of the galaxy interact in the regions where the young stars form. Recent observations of Centaurus A at radio and X-ray frequencies have greatly increased our knowledge of what is going on in this active galaxy. In one sense the new information has made Centaurus A more familiar. In another sense the observations have revealed physical processes so large and powerful that they make the active galaxy seem an even stranger and more wonderful phenomenon.

Centaurus A, which in observations at visible wavelengths is designated NGC 5128, was given its label because it is the most luminous source of radio waves in the constellation Centaurus. It occupies a position far south in the sky seen from the Northern Hemisphere. Seen from near Australia in the Southern Hemisphere, the constellation is overhead. Seen from New Mexico, where many of the recent radio observations have been made with the Very Large Array radio telescope (VLA) of the National Radio Astronomy Observatory, the galaxy is above the horizon only a few hours a day; its maximum elevation is about 14 degrees.

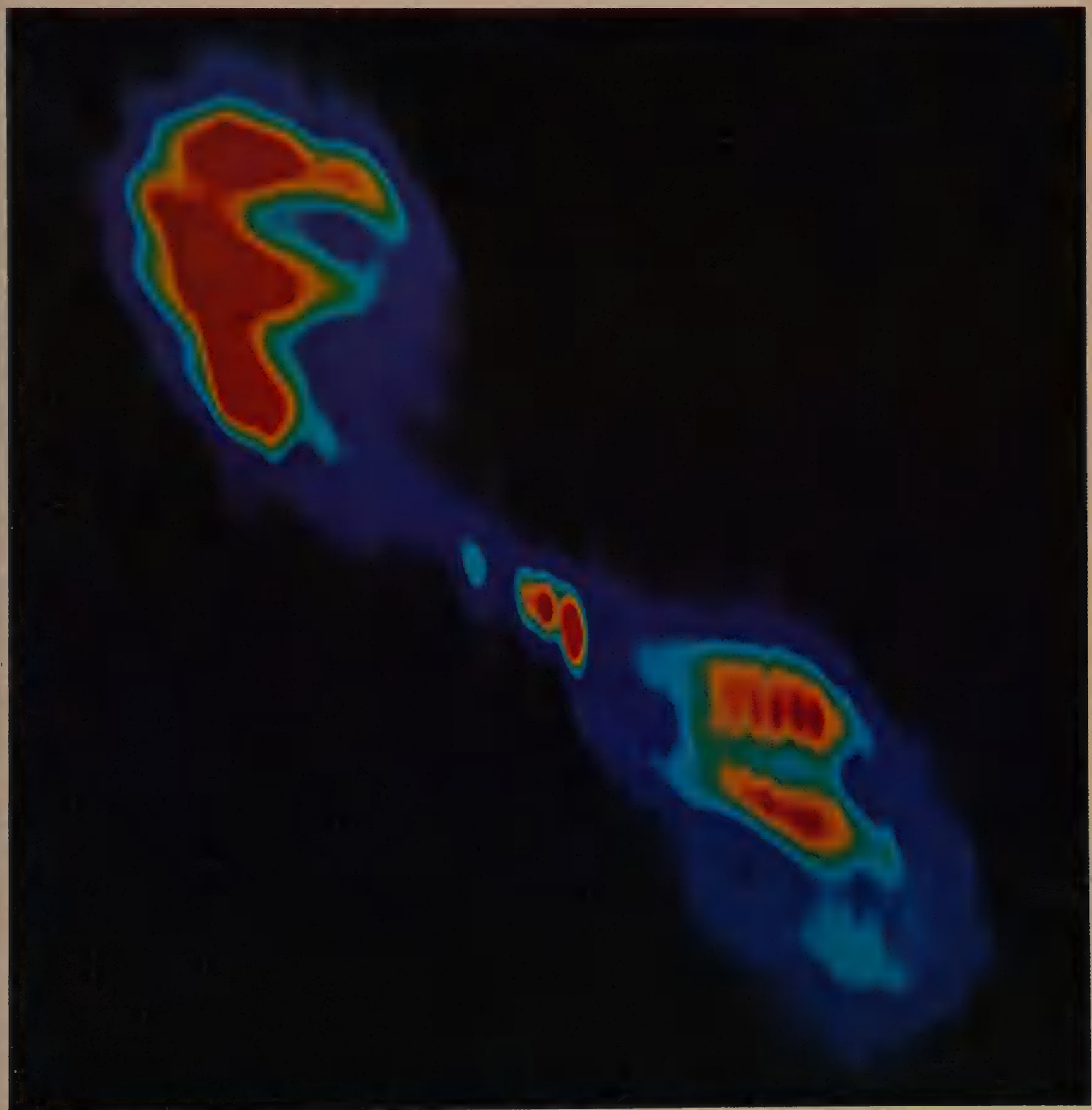
Centaurus A is one of the strongest sources of radio waves in the entire sky. The strength of the emission and the proximity of the object to our galaxy make it possible to observe processes and resolve structures in it that may be present in all active galaxies but that cannot be detected in more distant ones. As a result Centaurus A has had much attention from astronomers in the past few years. It has been studied with many of the most sensitive detectors of astronomical radiation, including in addition to the VLA the four-meter optical telescopes at the Cerro Tololo Inter-American Observatory in Chile and the Anglo-Australian Telescope at Siding Spring in Australia and the Einstein X-ray satellite observatory.

Much of the radiation detected in the work on Centaurus A originates in one of three physical processes. In all three processes radiation is emitted when an electron loses energy. The means by which the electron's energy changes, however, is different in each process, and the characteristics of the emitted radiation also differ. In the first process clouds of interstellar gas are energized by incident radiation. In such clouds atoms of gas absorb radiant energy and are thereby ionized (meaning that they gain or lose electrons). When an electron in the cloud comes near such an ion, the electron is deflected and accelerated; most of the ions are protons, hydrogen nuclei stripped of their single electron. As the electron's path shifts it emits a photon, or quantum of electromagnetic energy. Radiation generated in this way is referred to as thermal radiation because its spectrum depends strongly on the temperature of the gas.

Electromagnetic radiation can be described either as a stream of photons or

as a train of waves. In the photon description the radiation can be characterized by the energy of the photons; the corresponding property in the wave description is the frequency, which in turn is directly related to the wavelength. The greater the energy of the photon, the higher the frequency of the waves and the shorter the wavelength. In thermal radiation the energies of the emitted photons or the frequencies of the emitted waves are distributed continuously, that is, they form a smooth curve when they are plotted on a graph. The radiation tends to be most intense, however, in a fairly small range of frequencies. The waves are randomly polarized, meaning that the plane of the electromagnetic oscillations can have any orientation.

In the second process an electron is captured and accelerated in a helical path around a line of force in a magnetic field. As the electron is accelerated it



INNER RADIO LOBES of Centaurus A are shown on a false-color map of the radio emissions from the central region of the galaxy, which at visible wavelengths is designated NGC 5128. In the arbitrary color scheme red stands for the most intense radio emissions and blue for the least intense. Here and in other images of the galaxy on the following pages north is at the top and east is at the left. The inner lobes are the symmetrical rounded structures aligned on a northeast-southwest axis. The distance from tip to tip is about 60,000 light-years. The inner lobes are the smallest and innermost of a series of radio structures that extend from the center of the galaxy on each side. The narrow channel between the center of the galaxy and the northern inner lobe is called a jet. It is thought to be a beam of plasma, a highly ionized gas. The plasma is ejected from the galactic core, which is marked by the larger of the red ovals between the inner lobes. As in other active galaxies, the emissions from the plasma have a broad range of frequencies. The radio emissions were observed by one of the authors (Burns), Eric D. Feigelson and Ethan J. Schreier with the Very Large Array radio telescope (VLA) in New Mexico. The map was made by means of a computer image-processing system developed by Dennis Ghiglia of Sandia National Laboratories.

emits electromagnetic radiation tangent to the helical path. The emission constitutes what is called synchrotron radiation. Synchrotron radiation, like thermal radiation, has a continuous frequency distribution, but it is spread out over a broader range of frequencies than thermal radiation. Moreover, in synchrotron radiation the waves are polarized: the electric-field components of all the waves are aligned in the plane perpendicular to the magnetic line of force around which the electron spins.

In the third process a photon is emitted when an electron moves from one of the allowed energy levels in an atom to a lower level. When an atom in a cloud of gas absorbs energy from its environment, an electron in the atom can move to a higher energy level. Having ascended, the electron tends to quickly return to the original energy level. In the downward transition a photon with an energy equal to that of the photon absorbed in the upward transition is released. The photons emitted in this way have a set of discrete frequencies, each frequency corresponding to a particular transition within the atom. On a continuum representing the total electromagnetic spectrum the emitted radiation appears as a group of the bright lines known as atomic emission lines. The radiation in the emission lines is randomly polarized.

As we shall see, synchrotron radiation has turned out to be of great significance in understanding the physics of Centaurus A. The earliest observations of the galaxy, however, were made by means of radiation in the visible range, which is not surprising since NGC 5128 is the seventh most luminous object in the sky outside our own galaxy. From a wealth of optical observations beginning early in the 19th century it has become clear that the visible part of Centaurus A has a very unusual structure, with features of both elliptical and spiral galaxies.



OPTICAL IMAGE AND RADIO CONTOURS of Centaurus A are aligned perpendicular to each other. The illustration combines the radio contours of the inner lobes with a picture of the galaxy made with radiation in the visible range. The visible area of NGC 5128 has an unusual form with features of both elliptical and spiral galaxies. The large luminous circle is much like the body of an elliptical galaxy. The dark stripe that bisects the circle is a disk-shaped dust lane composed of gas, dust and associations of stars; the dust lane resembles a spiral galaxy. The disk is warped in opposite directions at its ends. Although the ellipse is rotating slowly if at all, the disk of the dust lane is rotating steadily. The radio structure emerges from the center of the dust lane more or less along its axis of rotation.

The main body of NGC 5128 is a bright ellipse about seven minutes of arc in angular extent. At the estimated distance of 15 million light-years each minute of arc is approximately 5,000 light-years, and hence the ellipse is about 35,000 light-years across at its widest. The mass of the galaxy has been estimated at about 300 billion times the mass of the sun. This mass and size are typical of elliptical systems.

What is unusual is the presence of a broad absorption lane, or lane of dust, in the middle of the ellipse. In optical images the dust lane appears as a dark stripe aligned with the short axis of the ellipse. The lane has the shape of a disk with its edges warped in opposite directions, like the brim of a fedora.

The dust lane is populated by dust, associations of stars and regions filled with ionized atomic hydrogen. The neutral, or un-ionized, hydrogen atom is designated H I and the ionized atom is designated H II. Therefore the regions of ionized hydrogen observed in the dust lane of NGC 5128 are referred to as H II regions. In the H II regions the hydrogen gas has a temperature of some 10,000 degrees Kelvin (degrees Celsius above absolute zero). The gas is heated by radiation from hot young blue stars that have recently formed in the gas cloud. The H II regions give off a set of characteristic atomic emission lines.

The main body of Centaurus A, like that of many elliptical systems, is rotating very slowly if at all. The dust lane, however, is rotating steadily and in a pattern typical of spiral or disk galaxies. The dust lane and the ellipse are centered at the same point, and the gravitational potential of the large ellipse dominates the system composed of the two components. Such a hybrid configuration is extremely rare and could well be connected with the violent activity underlying the radio structure.

Before turning to the radio and X-ray observations that are the center of current interest in Centaurus A, one other unusual optical feature is worth noting. Long-exposure photographic plates made by J. A. Graham of Cerro Tololo show there is a remarkable elongated luminous region in the northeast quadrant of the galaxy. The plates were made with the light of the atomic emission line corresponding to the recombination of an electron with a proton in the H II regions, which is referred to as the H-alpha transition. The plates show that at a distance of 50,000 light-years from the center of the galaxy there is a linear filament including three large H II regions aligned radially outward from the center of the galaxy. Farther out, at distances of up to 130,000 light-years, are regions yielding emission lines and including both filaments and diffuse gas. Spectroscopic studies show that the excitation of the gas decreases with distance from the galactic core, suggesting the source of the excited gas is in the center of the galaxy.

Near some of the elongated H II regions are what appear to be chains of bright blue young stars; the chains were first noted by Victor M. Blanco and his colleagues at Cerro Tololo. If these are indeed stars less than 10 million years old, they must have formed near where they are now observed rather than forming at the center of the galaxy and then being ejected. The stars are about 65,000 light-years from the galactic core, and even at the unusually high velocity of 1,000 kilometers per second (which is many times the velocity needed to escape the gravitational field of the main body of the galaxy) the stars could not have existed long enough to travel from the center of the galaxy to where they now are.

The first detailed radio observations of Centaurus A were begun in 1961 when the 210-foot radio telescope at the Australian National Radio Astronomy Observatory came into service. One of us (Price) was fortunate enough to work as a graduate student with J. G. Bolton and B. F. C. Cooper of that observatory in utilizing the new telescope, which was then the premier instrument in radio astronomy, to observe Centaurus A. It was quickly found that the radio emission from the galaxy is plane-polarized, suggesting it comes from the synchrotron process.

Further work showed that the plane of polarization of the emission is rotated as the radio waves pass through clouds of cosmic-ray electrons and the weak magnetic fields in the interstellar space of our galaxy. Such rotation, which is called the Faraday effect, is proportional to the square of the frequency of the radiation. Although the Faraday effect had been predicted theoretically, the observations of Centaurus A with the Australian radio telescope marked the first time the effect had been clearly detected observationally.

Thus the early radio investigation of Centaurus A had two significant results. First, the existence of the interstellar magnetic fields in our own galaxy was confirmed. Second, the confirmation of the hypothesis that the radio emissions from NGC 5128 come from a synchrotron process was a significant step in understanding the active galaxy.

Furthermore, the Australian radio observations provided the first reliable map of the overall radio structure of Centaurus A. It was found that the linear extent of the radio region is 2.7 million light-years, covering about 10 degrees in the sky. When Centaurus A was first mapped in the early 1960's, it was the largest discrete astronomical object that had ever been observed. In the past 10 years larger radio sources have been found, but Centaurus A is still among the largest astronomical objects known.

The extended radio-emitting regions on each side of the galactic center change their direction with the distance from the nucleus. Near the nucleus the alignment is northeast-southwest but the outer regions are aligned almost north and south. The shift in direction of about 65 degrees from the center of the galaxy to the tip of the radio envelope is caused by a continuous curvature along the length of the radio region. As a result the overall radio structure has the shape of a huge *S*. The curvature indicates that the flow of material into the radio envelope is being continuously perturbed. The perturbation could be due to a precession (a wobble like that of a top) of the nuclear engine or to the effect of gas in the galaxy outside the nucleus pressing on the plasma beam.

Between the galactic nucleus and the tip of the radio envelope is a series of complex radio structures. The size of these substructures increases from the nucleus outward, and the scale of the largest ones is about a million times that of the smallest. The details of the radio components have been resolved only recently in work with the VLA done by us, Eric D. Feigelson of Pennsylvania State University, Ethan J. Schreier of the Space Telescope Science Institute and George W. Clark of the Massachusetts Institute of Technology.

In a radio interferometer such as the VLA radio signals from a single source are received simultaneously by an array of several antennas. In the VLA there are 27 antennas arranged along three intersecting arms. Comparing the phase relations of the waves arriving at the receivers can give a rich picture of the radio source. With the VLA we have been able to achieve a resolution of one second of arc. At the distance of Centaurus A such a resolution is equal to about 80 light-years and is comparable to that in photographic plates made with visible radiation gathered by large telescopes.

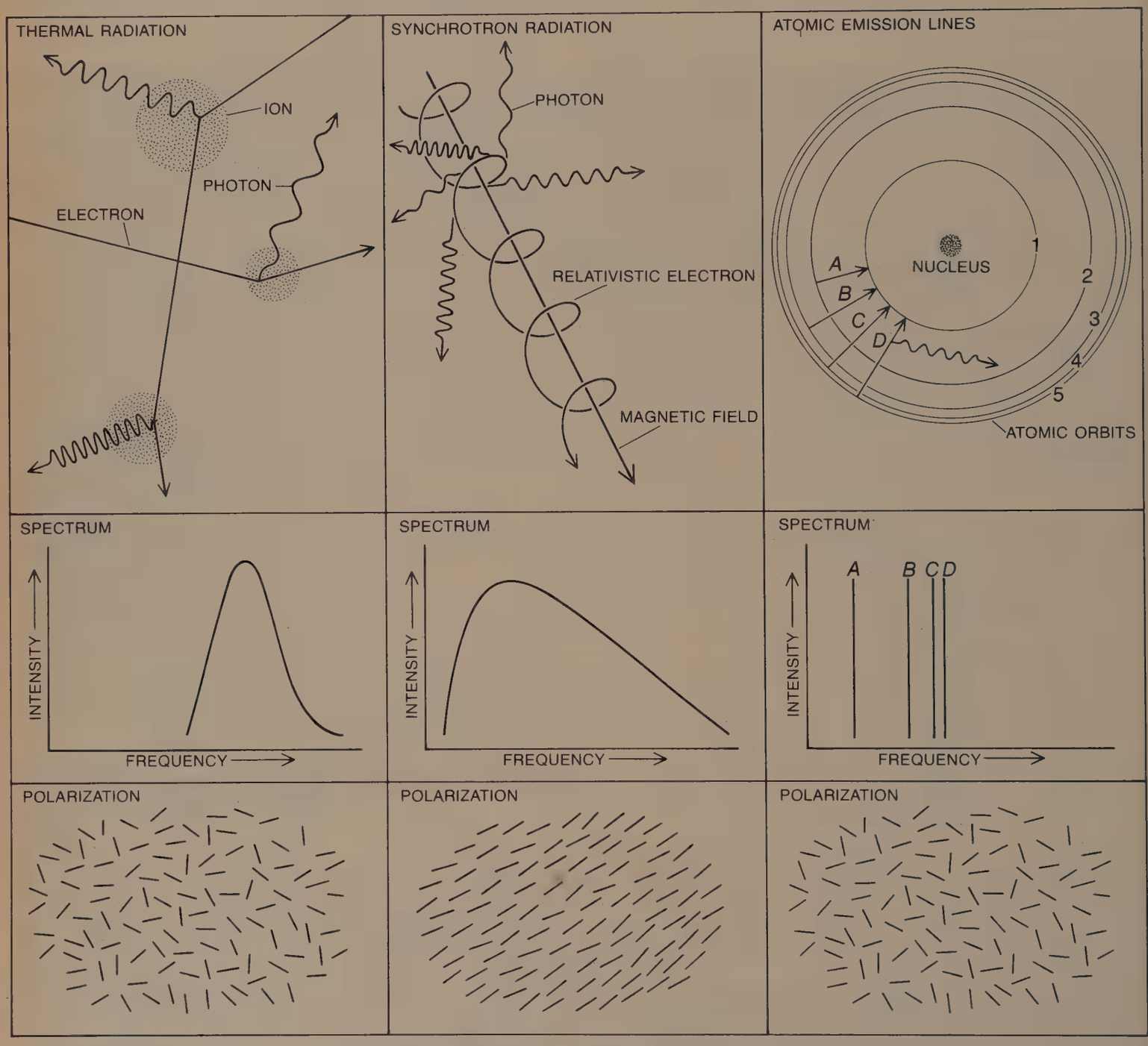
Moving inward toward the core from the ends of the radio structure an asymmetry becomes clear. To the northeast, between 70,000 and 130,000 light-years from the nucleus, is a large, roughly circular middle radio lobe. No comparable middle lobe is seen at the corresponding position in the southwest. The middle lobe is a region of intense emission that is distinct from the diffuse outer radio envelope. The position of the middle lobe coincides with that of the optical filaments where young stars are seen. As we shall see, it also overlaps a newly discovered area of X-ray emission.

Closer to the nucleus is a pair of inner radio lobes, one lobe on each side of the core. The inner lobes, which are also distinct from the outer envelope, are symmetrically positioned 30,000 light-years

from the center of the galaxy. With a diameter of 3,500 light-years they are considerably smaller than the middle lobe. No visible radiation or X-ray emission has yet been observed in association with the inner lobes.

Connecting the northern inner lobe and the nucleus is a narrowly collimated channel of radio and X-ray emission called a jet. Such jets, which are observed in many radio galaxies, could be channels that carry freshly energized electrons and magnetic fields from the central engine to the inner lobe. If this is so, the jet is a leaky and inefficient channel, since some of its energy is radiated outward as synchrotron emission.

Perhaps the most striking asymmetry in Centaurus A is the fact that only the northern part of the radio envelope has a jet; no comparable structure is observed in the southern part. Many radio galaxies have a jet extending from the galactic core, and in some of them the jet is one-sided. The reason for the asymmetry is not known, but several hypotheses have been proposed. Lawrence Rudnick of the University of Minnesota has suggested that periodically the jet switches sides. As a result over any long period the lobes would be supplied with roughly the same amount of energy. The oscillation could be due to the interaction of the jet with the dense gas in the region of the nucleus. Other workers have proposed that jets actually extend from both sides of the core but that because of relativistic effects the emission from the jet coming toward us is greatly enhanced and the emission from the jet going away from us is greatly reduced



RADIATION EMITTED BY CENTAURUS A could originate in one of three physical processes. In all three an electron loses energy, and the mechanism by which the energy is dissipated determines the frequency and intensity of the radiation. In thermal radiation (*left*) electrons are deflected and accelerated as they pass near ions (mostly hydrogen nuclei, or protons) in a cloud of hot gas. When an electron is accelerated, it gives off a photon, or quantum of radiation, which can be said to have an energy, a frequency and a wavelength. The frequencies of thermal radiation are continuously distributed, but the radiation is most intense in a fairly narrow range. The polarization of thermal radiation is random. In synchrotron radiation (*middle*) photons are given off by relativistic electrons: electrons with a velocity approaching that of light. The relativistic electrons are accelerated in a helical path around lines of force in a magnetic field and photons are emitted tangent to the path. Synchrotron radiation has a broad range of frequencies and is polarized. An electron can also give off radiation when it makes a downward transition from one allowed energy level in an atom to a lower level (*right*). The spectrum of radiation that is emitted in this way consists of a series of discrete lines. The radiation in these atomic emission lines is randomly polarized.

[see "Cosmic Jets," by Roger D. Blandford, Mitchell C. Begelman and Martin J. Rees; SCIENTIFIC AMERICAN, May, 1982]. This explanation, however, requires that the bulk speed of the plasma be close to that of light, which appears to be unlikely in Centaurus A.

In radio interferometry the resolution of the observations increases when the base line, or the distance between antennas, is increased. The resolution that can be achieved with a radio interferometer has recently been improved by putting antennas on different continents. The data from each instrument are recorded on a magnetic tape and are later correlated and compared by computer. Robert A. Preston and his colleagues at the Jet Propulsion Laboratory of the California Institute of Technology employed radio telescopes in Australia and South Africa to observe Centaurus A and were able to resolve features as small as three milliseconds of arc in the nucleus of the galaxy.

Deep in the core of Centaurus A, Preston and his colleagues found a jet about 50 milliseconds of arc long. At the distance of Centaurus A, 50 milliseconds of arc is equal to four light-years. The small jet emerges at the same angle as the larger radio jet found with the VLA. This remarkable collimation of the beam from four light-years out to 20,000 light-years means that the engine in the core maintains a consistent spatial orientation over long periods. The small inner jet is very close to the engine and is probably the initial point where the plasma beam becomes collimated.

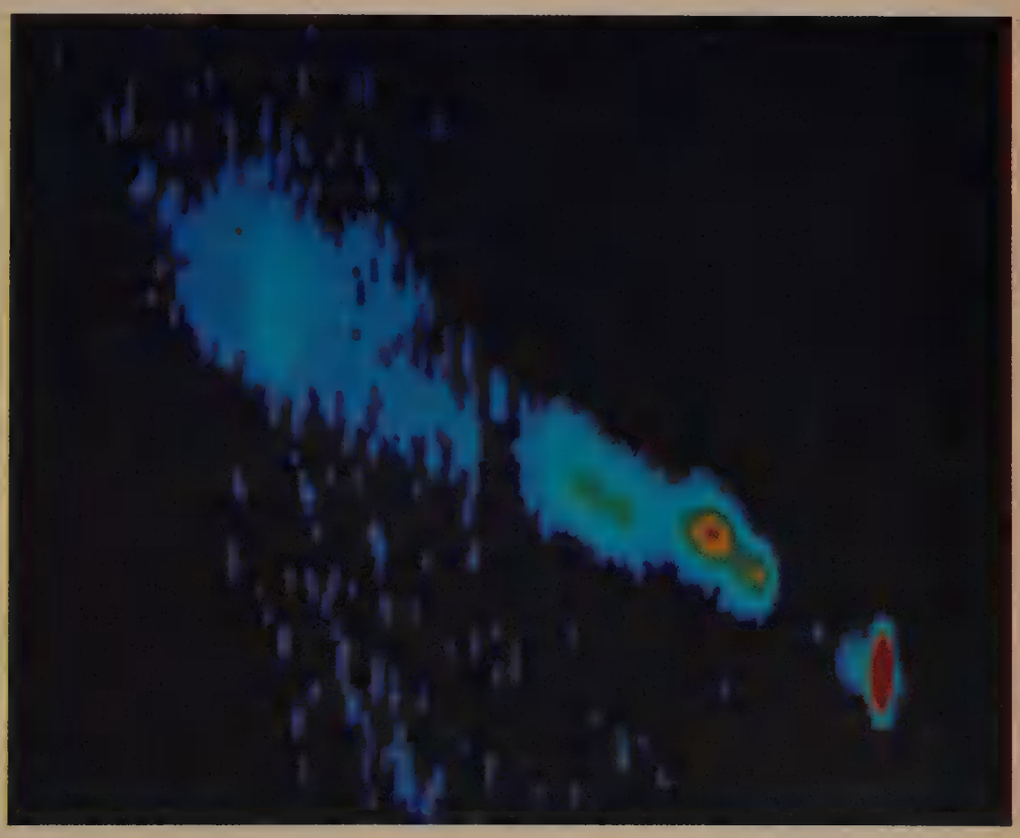
Completing the radio map of the galaxy is the core itself. In radio observations the core appears as a strong source of emissions at the very center of the galaxy; it cannot be resolved into its components with current instruments. While the detailed structure of Centaurus A was being examined by means of radio interferometers another significant kind of information about the galaxy was being accumulated.

The X-ray telescope in the Einstein observatory satellite, launched in November, 1978, was the first instrument capable of detecting and mapping soft X rays from sources other than the sun. Within the X-ray part of the electromagnetic spectrum soft X rays are those with a relatively low energy: from about 1,000 to 10,000 electron volts. Hard X rays correspond to photons with a high energy: from about 10,000 to 50,000 electron volts. The two main detectors mounted on the Einstein satellite were able to resolve X-ray features respectively 1.5 minutes of arc and five seconds of arc across, for the first time giving X-ray astronomers a resolving power comparable to that of radio telescopes and optical telescopes.

Soon after the Einstein satellite was



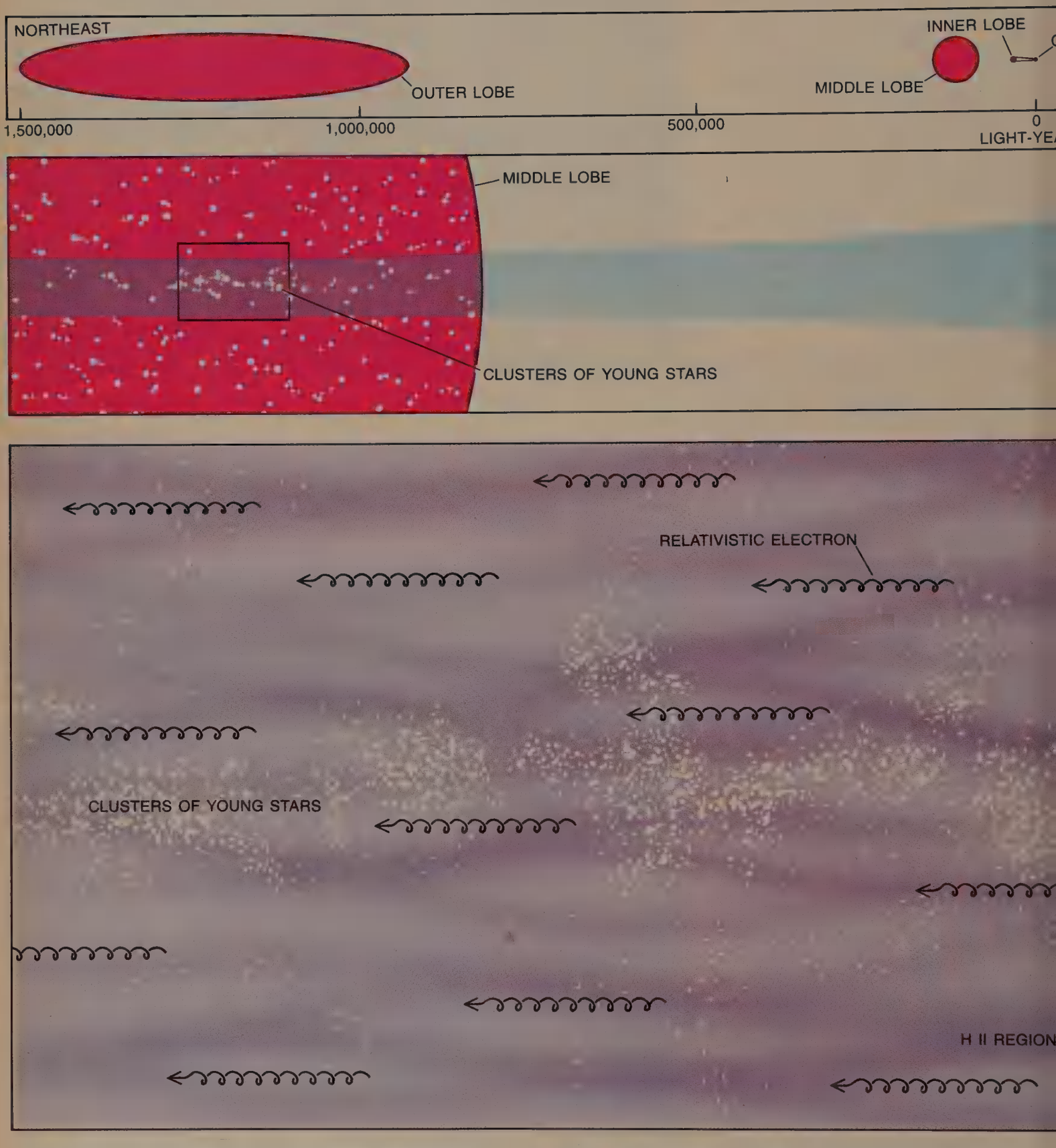
INNER JET is a highly collimated beam of plasma that channels freshly energized electrons from the core of Centaurus A to the northern inner lobe. The galactic nucleus is the compact red region at the lower right with white contours superposed on it; the jet extends to the northeast from the core. The jet is about 3,000 light-years long. The computer image combines radio data with X-ray data gathered by Feigelson and his colleagues using the Einstein satellite observatory. The colored areas correspond to X-ray emission, the white contours to radio emission. The conjunction of X-ray and radio emissions suggests that the radiation from the plasma comes from a synchrotron process. No corresponding jet is seen to the southwest of the core.



HIGH-RESOLUTION IMAGE of the inner jet is a detail of the illustration at the top of the page: it expands a small area near the center of the galaxy. The galactic core is the red oval at the lower right. The colored areas to the northeast of the core are "knots," regions of particularly intense radio emission. Since the relativistic electrons in the plasma can retain their high energy for only about 50 years, they must be reaccelerated in order to reach the inner lobe, which is 3,000 light-years from the core. The knots could be turbulent regions in the jet where the electrons are reenergized. The image was made utilizing the VLA in a configuration that makes it possible to resolve features as small as 80 light-years across at the distance of Centaurus A.

launched Feigelson and Schreier, together with Riccardo Giacconi of the Center for Astrophysics of the Harvard College Observatory and the Smithsonian Astrophysical Observatory, directed the satellite's instruments at Centaurus A and found a wealth of detail unprecedented in extragalactic X-ray astronomy. At about the same distance and angle from the nucleus as the northern middle radio lobe is a region of X-ray emission 24,000 light-years long. The radiation is more intense than would be expected from synchrotron emission. Therefore it is probably thermal radiation; its source could be a cloud of gas with a temperature of 10 million degrees K. The gas might be heated when it is compressed by turbulence in the radio plasma of the middle lobe.

An envelope of less intense X-ray

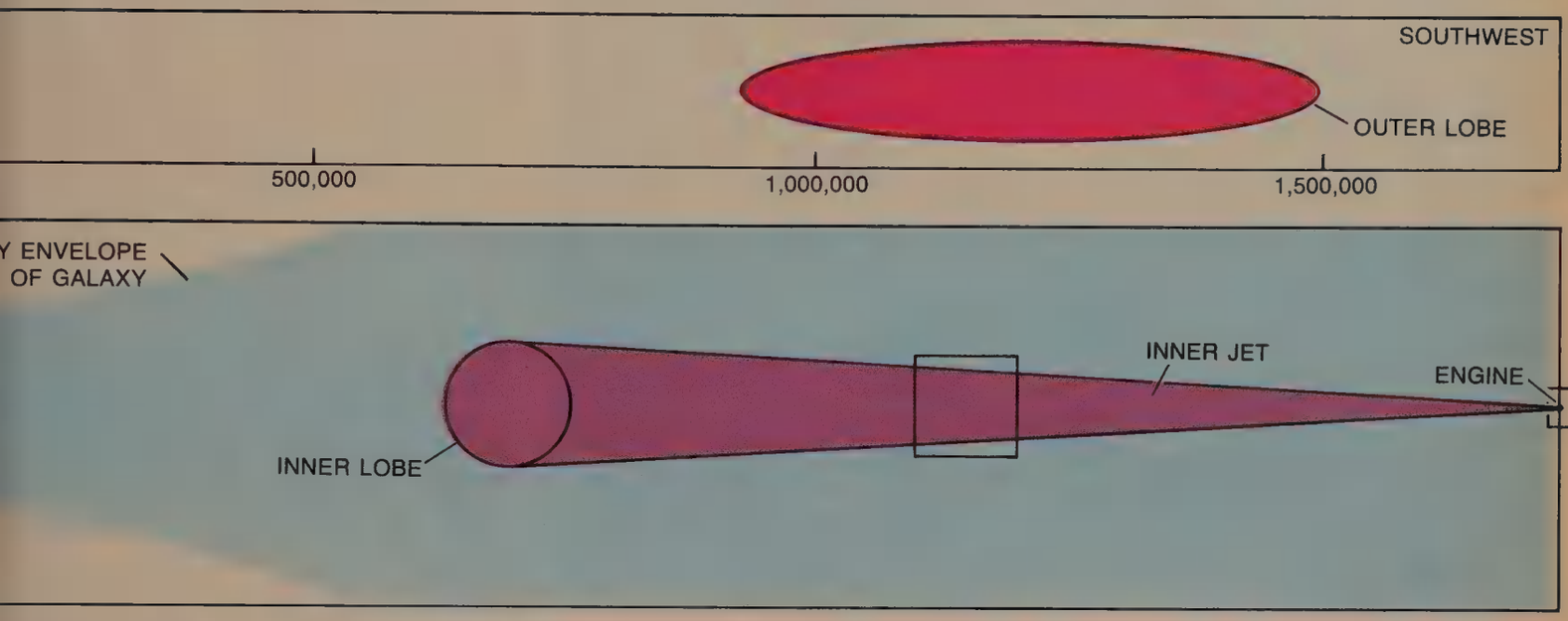


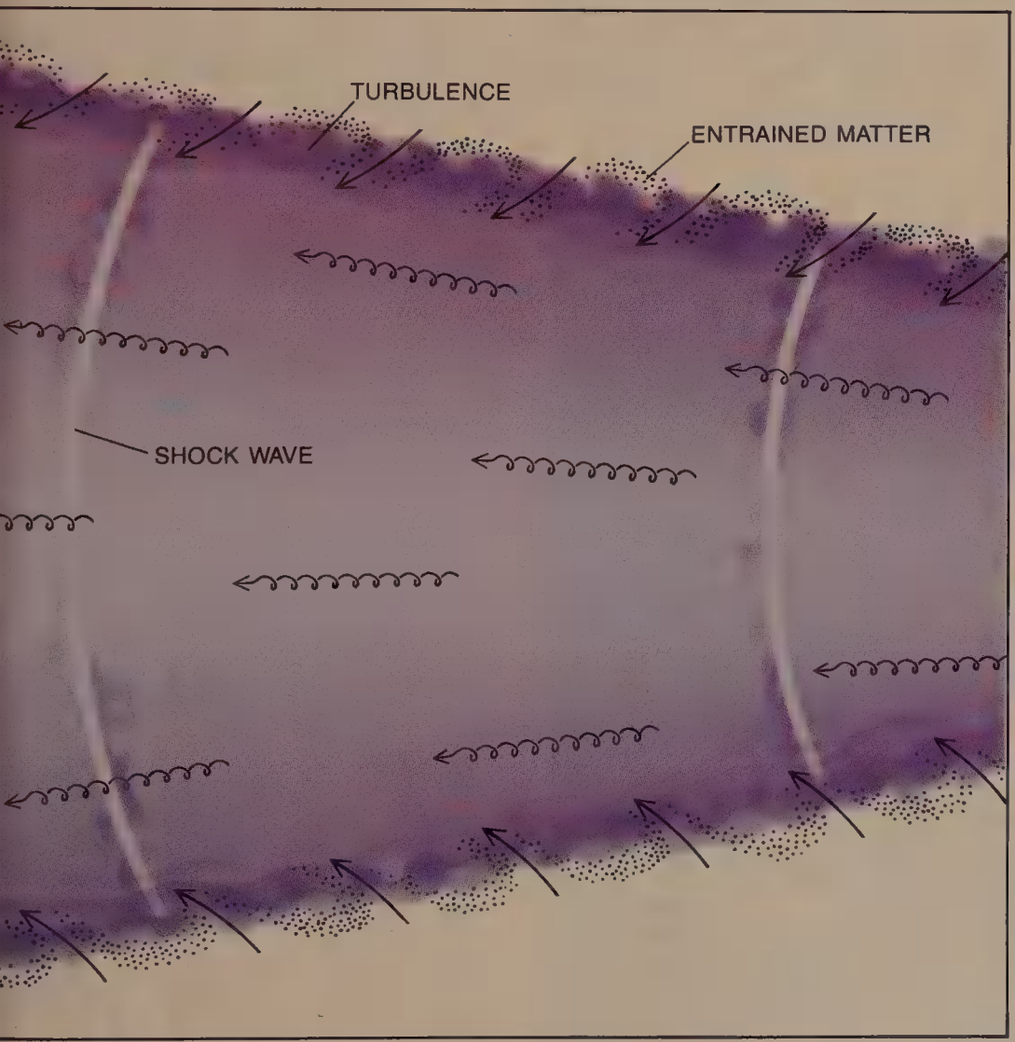
DYNAMICS OF THE EMITTING REGION of Centaurus A are shown schematically in panels with a progressively smaller scale. The full extent of the emitting region is three million light-years (*top panel*). Radio emissions are shown in red and X-ray emissions in blue. The pair of inner lobes on each side of the center of the galaxy are areas of relatively intense radio emission within a diffuse outer emitting envelope. The inner jet that connects the core to the northern inner lobe gives off much emission at both radio and X-ray wavelengths (*middle panel*). It lies within a larger region of diffuse X-ray emission that comes from the interstellar medium. The region of diffuse X-ray emission extends to the northern middle lobe. One edge of the middle lobe is also the site of a group of extended filaments and bright blue point sources of visible radiation. One hypothesis that accounts for many recent radio and X-ray observations of Centaurus

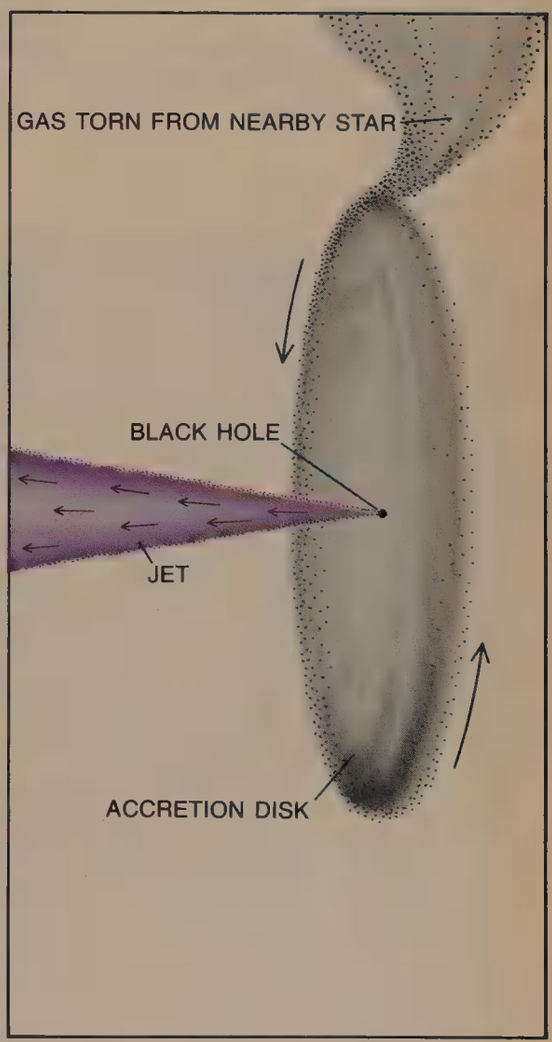
emission surrounds the galaxy, extending out to a distance of about 10,000 light-years from the core. The diffuse emission is probably generated by a cloud of interstellar gas with a temperature of 20 million degrees K. and a total mass 200 million times that of the sun.

The detection of the hot gas by the instruments on board the Einstein satellite is one of the few times an interstellar medium has been found within the visible body of an elliptical galaxy.

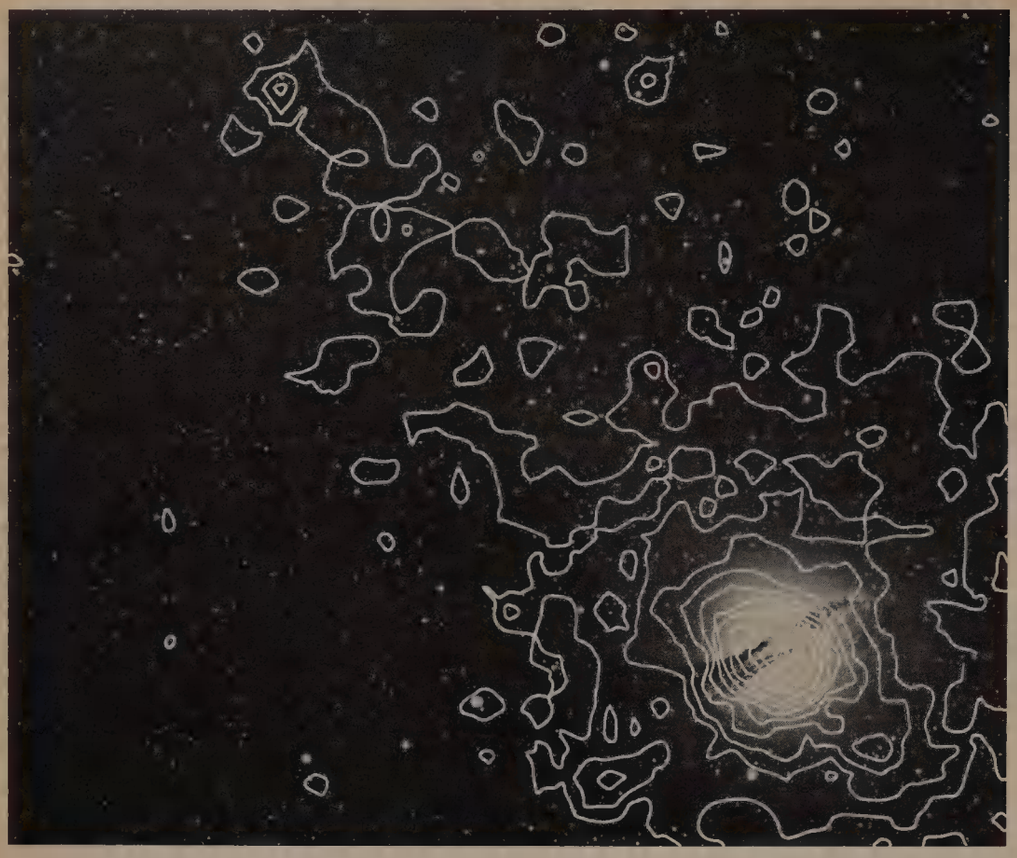
With a density of about one particle per 300 cubic centimeters the interstellar medium discovered in the Einstein satellite observations is fairly dense by astronomical standards. The dense medium could confine the plasma in the radio jet and the inner radio lobe and thereby contribute to the maintaining of their shape. Such confinement could be



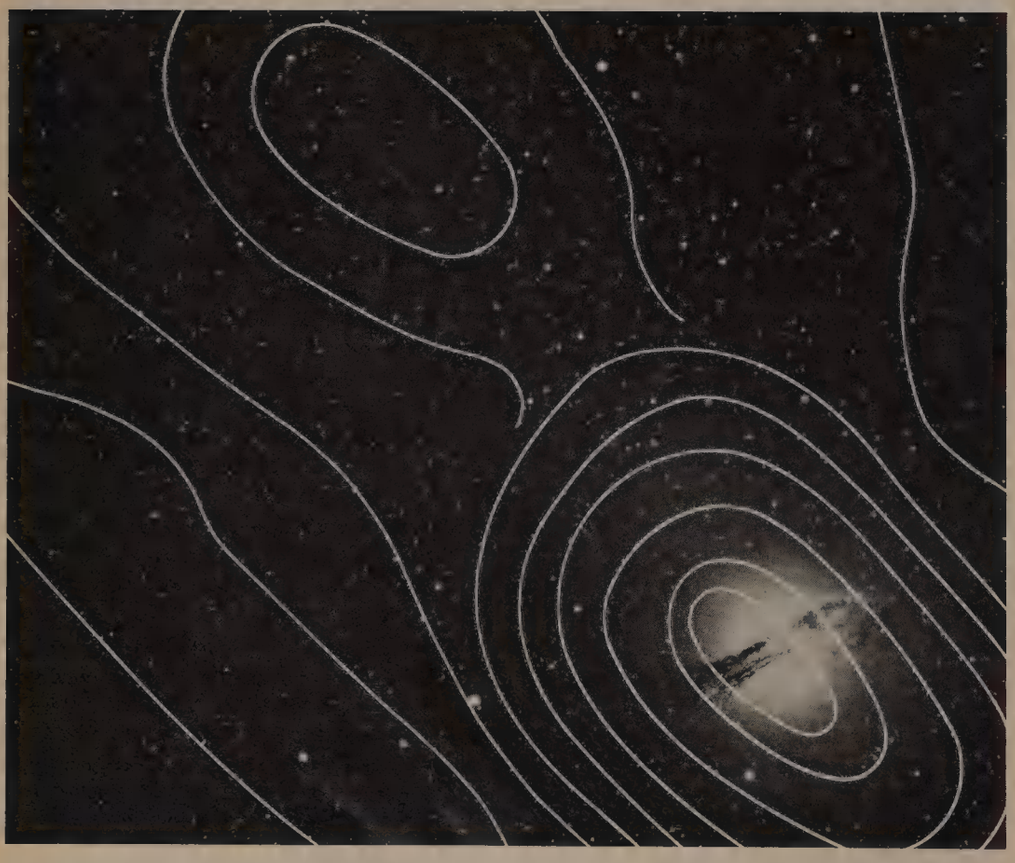




A is the following: The engine that generates the radio plasma has at its center a black hole with the mass of a billion suns (*bottom right*). The black hole is surrounded by a toroidal accretion disk made up of gas and dust. Fuel for the engine comes from infalling gas, which could be tidally torn from the atmosphere of stars near the disk. The interaction of the gas with the black hole and the accretion disk yields a narrowly collimated beam of electrons and magnetic fields: the inner jet. As the jet moves outward from the core, turbulence at the edge of the jet causes gas to be picked up and carried along with the plasma (*bottom middle*). Within the jet shock waves reenergize the relativistic electrons. When the plasma reaches the region of the middle lobe, some of the entrained gas cools and contracts and forms stars (*bottom left*). At the end of their life they explode as supernovas, putting energy back into the radio plasma and energizing the electrons.



X-RAY CONTOURS mapped with the Einstein satellite observatory are superposed on an image of NGC 5128 made by means of visible radiation. The main body of the galaxy is at the lower right. The optical filaments and young stars near the northern middle lobe are at the upper left; they are 130,000 light-years from the core. The chief X-ray emissions outside the main body of the galaxy coincide with the optical filaments. The optical image is made by the emission line resulting from the recombination of an electron and a proton to form a hydrogen atom.



RADIO CONTOURS made with the 210-foot radio telescope at the Australian Radio Astronomy Observatory are superposed on the same optical image of NGC 5128 in the illustration at the top of the page. As in the case of the X-ray emissions, the region of the most intense radio emission outside the main body of the galaxy (the northern middle radio lobe) coincides with the optical filaments and young stars. The overlapping of the sources of radiation at X-ray, visible and radio frequencies suggests that the same process could be responsible for all three.

critical. Without it the radio jet might lose its collimation and the lobes would expand greatly, diluting their energy density. If the hypothesis is correct, the structure of NGC 5128 confirms the thermal confinement of radio structures, a phenomenon that had been predicted but had never been observed.

The most exciting finding made with the Einstein satellite, however, was the discovery of a jet of X-ray emission coming from the core in much the same position as the VLA radio jet. This was the first time such an intense elongated structure had been found in X-ray data. Indeed, the X-ray jet was discovered before the VLA jet, and the X-ray findings motivated much of the recent work at radio frequencies.

Not only is the X-ray jet the same size and shape as the radio jet but also at the same places in the two jets there are "knots": small regions where the emissions are particularly intense. The structural similarities, along with the continuous spectrum of the emissions and their polarization properties, led us to conclude that the X-ray emission from Centaurus A is also synchrotron radiation.

When the data from more than two years of operating the Einstein satellite were analyzed, only one other galactic X-ray jet like the one in NGC 5128 was found. The second jet is in the next-closest active galaxy: M87. This galaxy is about three times as far from us as Centaurus A is. The resolution of the satellite's instruments, however, is not sufficient to show structural details in the M87 jet. The X-ray jet in Centaurus A has details that can be observed individually, and it therefore currently remains a unique astronomical object.

What does the detailed map of NGC 5128 that has been patiently assembled at X-ray and radio frequencies tell us about the physical processes going on in the galaxy? Consider the inner radio and X-ray jet. From the point of view of astronomical observation the most significant component of the jet is the stream of relativistic electrons (electrons moving at nearly the speed of light). The electrons are the source of the synchrotron radiation. It does not follow, however, that because the electrons are moving at relativistic velocities the jet as a whole has such a velocity. The beam includes a considerable quantity of gas in addition to the electrons, and the gas has a much lower speed than the electrons do. The best current estimate is that the bulk-flow velocity of the jet is about 5,000 kilometers per second.

The conclusion that the X-ray photons detected by the Einstein satellite come from a synchrotron process puts some severe constraints on the model of what goes on in the jet. The X-ray photons have an energy of about 2,000 elec-

tron volts. In the synchrotron process such photons are emitted by relativistic electrons with an energy of 3×10^{13} electron volts. Electrons with that much energy cannot stay in a magnetic field for long before they emit photons.

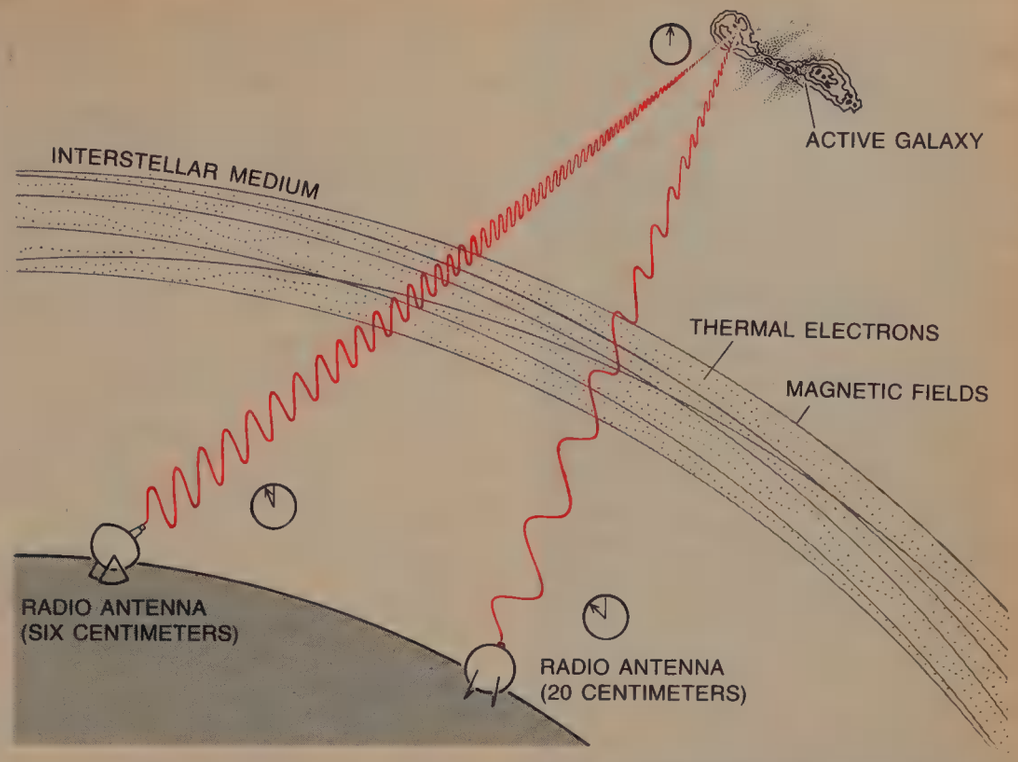
If the emission continues, eventually the electron will be depleted of energy and will not be able to emit any more photons. Indeed, given the energies of the electrons and the X-ray photons the depletion will occur in 50 years or less. The radio and X-ray jet in Centaurus A, however, reaches the inner lobe, which is 20,000 light-years from the nucleus. Therefore if the electrons are traveling at about the speed of light, they will have to be reenergized many times in order to reach the inner lobe.

The relativistic electrons in the beam could be accelerated by any one of three mechanisms. The first mechanism entails the existence of shock waves in the jet much like the ones formed in the earth's atmosphere when an airplane is flying at a velocity greater than that of sound. Such shocks could result from the collision of the jet with interstellar clouds. They could also form internally if the beam is unsteady or unstable. The knots in the beam would then correspond to regions where the plasma is subjected to strong shocks.

In the second mechanism the energy of turbulence in the beam provides the energy to accelerate the electrons. Instabilities and turbulence can form on the boundary layer between a beam of plasma and the medium that surrounds it. The turbulence could propagate into the beam in the form of waves whose motion would be damped by interaction with the electrons. In interacting with the waves the electrons would be accelerated. The net result of the process would be to transfer the bulk-flow energy of the plasma into the kinetic energy of the waves and then into the acceleration of the relativistic electrons.

The third possibility is that the electrons could be reenergized by collisions with the protons in the beam. The protons in the plasma cannot be detected directly because they do not emit a detectable quantity of radiation. For the same reason, however, they could retain their high kinetic energy much longer than the electrons do. Collisions between protons and electrons would eventually deplete the reservoir of proton energy, but the process would be slow and the reservoir might be large enough to get the plasma all the way to the inner radio lobe.

The process by which the beam of electrons, protons and magnetic fields gets from the inner lobe to the middle lobe is not yet well understood. However it operates, when the plasma reaches the middle lobe, it could have an important role in the formation of the young stars and the emission-line regions observed at distances of about 130,000 light-years from the core.



FARADAY EFFECT rotates the plane of polarization of radiation traveling through our galaxy. Radiation from an active galaxy such as Centaurus A is often plane-polarized because it originates in the synchrotron process. As the radiation passes through the interstellar medium in our galaxy it interacts with thermal electrons and magnetic fields. As a result the plane of polarization of the waves is rotated. The degree of rotation is proportional to the wavelength: the longer the wavelength, the greater the rotation. Radio observations of Centaurus A that were made in the early 1960's provided the first direct confirmation of the Faraday effect in the Milky Way galaxy. The observations also helped to strengthen the conclusion that the broad-spectrum radiation emitted by Centaurus A has its origin in a synchrotron process.

David S. De Young of the Kitt Peak National Observatory proposes the following scenario for the interaction of the radio structures and the optical structures. Eddies in the turbulent layer at the edge of the jet can entrain nearby material, which is mainly gas and dust. As the material is entrained it is heated by shocks in the plasma beam to about 10 million degrees K.

If the elements in the gas surrounding the jet have the same abundances as they do in the sun, which is a reasonable assumption for an evolved galaxy, the gas will cool to 10,000 degrees K. in from 10 million to 100 million years. At a temperature of 10,000 degrees the gas would show optical emission lines like those seen in the northeast quadrant of NGC 5128. Moreover, in the time it took the gas to cool it would have traveled between 33,000 and 330,000 light-years, a range that could put it near the position of the optical emission seen in NGC 5128 and in several other radio galaxies where optical emission lines coincide with radio lobes.

A fraction of the entrained gas would continue to cool and would eventually reach temperatures much lower than 10,000 degrees. The gas would then begin to coalesce by gravitational attraction and stars would form in it. De Young holds that the clusters of optical emission observed in the northeast quadrant of NGC 5128 support the conclusion that stars are forming in the region of the middle lobe.

The plasma contributes to the formation of stars by entraining ambient matter, but the stars could also help to keep the plasma beam energized as it moves outward from the center of the galaxy. Some of the stars formed in the northeast quadrant would probably have masses greater than 10 times the mass of the sun. Such massive stars have a quite short lifetime: about 10 million years. At the end of their relatively short life the massive stars would explode as supernovas.

The effect of the explosion would be to reenergize the electrons in the beam. Therefore the flow in the northeast quadrant could be self-perpetuating: the plasma beam gives rise to the formation of giant stars, which in turn explode and return their energy to the beam.

This account of how the plasma beam could be energized and how the optical filaments form is based on substantial observational data. Two of the most fundamental questions about Centaurus A and other active galaxies, however, must be answered in large part on the basis of conjecture. The questions are: What is the basis of the galactic activity? How does the central engine work?

It is possible that the activity observed

in NGC 5128 is directly related to the galaxy's unusual structure, which combines an ellipse with a disk. Some 30 years ago Walter Baade and Rudolph Minkowski of the Mount Wilson and Palomar Observatories hypothesized that the form of NGC 5128 is the result of a merger between a spiral galaxy and an elliptical one. Recent theoretical work by Allan D. Tubbs of the National Radio Astronomy Observatory has shown that gas and dust tidally torn from a nearby galaxy can fall into an elliptical galaxy and settle in a shape much like that of the dust lane in NGC 5128.

Such a merging process has been called galactic cannibalism. Cannibalism has been invoked to explain some powerful radio sources other than Centaurus A. Certain clusters of galaxies that include a giant galaxy at their center are five times as likely to be a strong radio source as the average cluster of galaxies. It has been proposed that the giant central galaxy is formed by repeated gravitational encounters deep in the cluster. The encounters generate a frictional force that causes the larger galaxies to spiral slowly inward toward the center of the cluster. Over the lifetime of the cluster the most massive galaxies continue to coalesce, ultimately forming a single massive central unit. The central giant galaxy then cannibalizes the smaller galaxies nearby.

The cause of the radio activity of such giant galaxies is almost certainly the matter accumulated in the process of cannibalism. The accreting gas, dust and stars provide the fuel for the radio engine at the core of the galaxy. A more limited cannibalism in NGC 5128 could also have triggered violent activity in the galactic nucleus and the corresponding radio activity. One problem with the cannibalism hypothesis is that Centaurus A is quite isolated: its only companions within a million light-years are a few dwarf galaxies. Perhaps the galaxy has cannibalized all its neighbors and the nonthermal radio and X-ray emissions are signs of cosmic indigestion in the aftermath of a hearty meal.

How might the central engine utilize the fuel the cannibalism provides? The fuel could probably be consumed most efficiently in the form of a cloud of gas. Observing with the VLA, Jan M. van der Hulst of the Westerbork Observatory in the Netherlands recently found clouds of neutral hydrogen gas within about 500 light-years of the core in Centaurus A. The pattern of the clouds' motion suggests that they are falling toward the nucleus. Such clouds could each year provide the engine with fuel equivalent to at least a tenth of the mass of the sun. If we assume that the engine has an efficiency of 10 percent in converting mass into energy, a tenth of a solar mass per year would be more than enough to account for the output of the engine, namely the extended radio region.

Fuel for the engine might also come from the atmosphere of stars in the central region of the galaxy. Jack G. Hills of the Los Alamos National Laboratory and others have suggested that stars near the core are subject to large tidal forces, implying that the strength of the gravitational field varies considerably from one side of the star to the other. As a result the star could be broken up or at least stripped of its atmosphere.

Hence matter is available as fuel for the engine at the core of the galaxy. The next problem is to find out how the clouds of gas, which would tend to circle the nucleus, are drawn into it. One way would be for two clouds of gas to collide in the vicinity of the nucleus. In such a collision one cloud might lose angular momentum and fall into the nucleus while the other gained momentum and moved outward.

Having accounted for the fuel and suggested a feeder mechanism whereby the fuel could be delivered to the engine, it is of great interest to consider what is going on in the engine itself. To do so it is useful to have an estimate of the size of the machine. The size of a source of radiation can be estimated by measuring the period over which the intensity of the emission varies. The reasoning underlying such an estimate is as follows. For two widely separated regions of the source to turn on or off simultaneously they must have some means of communication: a physical signal must pass between them. Since the fastest such a signal can propagate is the speed of light, it follows that if a significant fraction of the emission varies in, say, two hours, the source itself cannot be more than two light-hours across.

Although this reasoning is not incontrovertible, it is probably valid for a source with a continuous emitting surface, and we have employed it to estimate the size of the central engine of NGC 5128. Observations of the variability of the X-ray and radio fluxes from the core of Centaurus A show that the emissions vary significantly over a period of months. Faster variations, however, are also observed, with the fastest taking less than 24 hours. Therefore we conclude that the core has components ranging in size from light-months to light-days.

Although any account of what actually goes on in the substructures of the galactic engine is highly speculative, one point seems clear. Whatever the engine's mechanism is, it is not nucleosynthesis, the process of nuclear fusion that supplies the power for most common stars. The assumption of an efficiency of 10 percent for the central engine rules out nucleosynthesis, which has an efficiency of less than 1 percent and would thus require more fuel and more time to generate the radio structure of Centaurus A than is suggested by recent observations. Moreover, nucleosynthesis yields thermal radiation rather than the nonthermal synchrotron emissions observed in NGC 5128.

A more efficient way for energy to be extracted from matter is for the matter to fall into a strong gravitational field. When that happens, the matter gains considerable kinetic energy. If the energized fuel then collides with structures in the core, energy can be released in the form of high-energy electromagnetic waves and even high-speed particles.

For this to be the mechanism of the central engine of Centaurus A there would have to be a source of a very strong gravitational field at the center of the galaxy. Such a field could be provided by what is referred to as a collapsed object: a black hole with a mass of about a billion solar masses. If the collapsed object exists at the center of the galaxy, it is undoubtedly spinning, since it is difficult to imagine a process that would lead to the creation of a black hole without at the same time giving it considerable angular momentum.

The spinning collapsed object would have three main effects. First, it would provide the well of gravitational potential into which the fuel could fall. Second, its spin axis would orient the entire engine. The spin would cause nearby matter to precess around the engine and form the disk of gas known as an accretion disk. Third, the black hole could have a magnetic field associated with it, and the lines of force in the field could accelerate charged particles to a high energy, extracting energy from the black hole as they did so.

Imagine a cloud of relatively cold gas falling onto the accretion disk. The disk might have the shape of a torus with the black hole in the center. Some of the infalling gas would be collimated within the narrow confines of the black hole and driven back out along the rotation axis of the hole by the radiation pressure of emission from the inner surface of the accretion disk. In this way the disk could both collimate and accelerate the particle beam that forms the inner radio jet.

For the moment this is as far as even the boldest speculation can take us in understanding active galaxies such as Centaurus A. The galaxy remains the focus of intense scientific activity. Because of its proximity and its intrinsic interest, it will undoubtedly continue to be so over the next few years. The current work has amassed much information on the structure and physical processes of Centaurus A. The most intriguing aspect of an active galaxy, however, is the mechanism that underlies its huge emitting region. We believe the study of Centaurus A will help to solve this fundamental problem in astrophysics.



© 1983 SFES

among a select number of people guaranteed the opportunity to own every one of these Signed First Editions.

Major works by great authors

The Society has already arranged to obtain and review new manuscripts by such prominent authors as Allen Drury (Advise and Consent), John Hersey (The Wall and A Bell for Adano), William Styron (The Confessions of Nat Turner and Sophie's Choice), Wallace Stegner (The Spectator Bird), Leon Uris (Exodus, Armageddon, and Trinity).

Indeed, the authors to be published by the Society in the coming months have accounted for eight Pulitzers and six National Book Awards. There will be a searing new novel by Philip Roth. A fascinating historical study by Louis Auchincloss. A biographical novel by Irving Stone. And eagerly awaited new books by such best-selling authors as Irwin Shaw, Joyce Carol Oates, Walker Percy, Gore Vidal, Arthur Schlesinger Jr., Robert Penn Warren and Kurt Vonnegut Jr.

The First Editions of all of these new books will be reserved exclusively for members of the Society.

The value of owning author-signed First Editions

Among knowledgeable book collectors, First Editions have traditionally been held in highest esteem. Especially prized are First Editions that have been signed by their authors. For then the book—like a signed work of art—serves as a direct link between the owner and the writer.

Accordingly, as each new book is prepared for publication, every aspect of the volume—from the marbled endleaves of the book to the unique design of its leather binding—will be carried out to reflect the spirit and intent of the work. And the author's hand-written signature in the book will be a final mark of approval.

In addition, each author will write a special foreword that will appear only in the Signed First Edition issued by the Society—adding to the importance of owning these superb volumes.

As a charter member, therefore, you will possess perhaps the most desirable books of all: important new First Editions, bound in leather, and signed for you by the greatest authors of today. A private library to be treasured throughout a lifetime—and by generations to come.

Leather accented with 22kt gold

Every one of these Signed First Editions will be created to enhance the collector's library and the reader's enjoyment.

Each volume will be fully bound in carefully selected hand-cut leather, deeply stamped with an original cover design. The leather of the spine will be "hubbed" with raised horizontal ridges, and richly decorated in 22 karat gold. And throughout the collection, the books will vary in height, size, cover design, and in the color and grain of their leather bindings.

Leading artists and book illustrators will be commissioned to create original works of art exclusively for these volumes—they will never appear in any other editions.

Further, only quality bookpapers will be used—custom-milled to be free of the acids that cause ordinary papers to yellow. Pages will be thread-sewn into separate "signatures" and the edges gilded with a tarnish-free finish to seal the volume from dust and moisture. Finally, a ribbon page-marker will be bound into each book—an elegant touch and added convenience.

These will be volumes of such beauty and craftsmanship that their authors will be proud to sign them. And you will be equally proud to own them.

Long-term price guarantee—at no obligation

The total number of each new First Edition will be limited to just one volume per member. No additional copies will ever be made available.

By enrolling now, you will receive each new Signed First Edition—approximately one book each month—at the original issue price of just $35. This price will be guaranteed to you for three full years. And you will have the right to cancel your membership at any time, simply by giving thirty days' written notice.

Application deadline: December 31st

To enroll as a charter member, simply complete and return the accompanying application by December 31, 1983. After that date, the charter membership rolls will be permanently closed.

In future years, the Society may accept non-charter members, who will be able to acquire only those volumes made available during the periods of their membership—at prices then prevailing.

To be sure of acceptance now, mail your Charter Membership Application by the deadline of December 31, 1983. No payment is required now, but it is important to respond by that date.

THE FRANKLIN LIBRARY
A FRANKLIN MINT COMPANY



CHARTER MEMBERSHIP APPLICATION

The Signed First Edition Society

Must be postmarked by December 31, 1983.

The Franklin Library
Franklin Center, Pennsylvania 19091

Please enroll me as a Charter Member of The Signed First Edition Society, and send me the First Editions to be published by the Society during the period of my three-year membership. Each volume will be fully bound in leather, the spine accented with 22 karat gold, and personally signed by the author. I understand there will be approximately 12 books published each year, and that I can cancel my membership at any time upon 30 days' written notice.

I need send no money now. I will be billed $35.* for each leather-bound volume in advance of shipment. This price will be guaranteed to me as a Charter Member for three full years.

*Plus my state sales tax and $2.50 for shipping and handling.

Signature _____
ALL APPLICATIONS ARE SUBJECT TO ACCEPTANCE.

Mr.
Mrs.
Miss _____
PLEASE PRINT CLEARLY

Address _____

City _____

State, Zip _____

Limit: One Membership per person.

Canadian residents will be billed for each book in advance of shipment at $47. plus $2.50 (Cdn. $).

SCIENCE AND THE CITIZEN

Lost Ground

Although the U.S. is in principle an egalitarian society, it is well known that some social groups have faced greater barriers to social and economic achievement than others. The group that has been confronted by the most substantial obstacles consists of the black citizens of the U.S. Many efforts have been made to remove such obstacles. Have they succeeded in bringing blacks into any kind of parity with whites? According to a recent report by the Bureau of the Census, the answer is clearly no. Indeed, the report suggests that by several significant measures of social welfare, including average income and the integrity of the family, between 1970 and 1982 blacks lost ground to whites.

The report was compiled from several sources, among them the censuses of 1970 and 1980, monthly surveys of statistically representative samples of the U.S. population done between censuses by the bureau and reports from the Bureau of Labor Statistics, the National Center for Health Statistics and the Department of Defense. The accumulated data show that blacks living in an intact family are doing better economically in relation to whites than they were a decade ago. Families without a husband present, however, are doing worse. Moreover, the proportion of households headed by a woman is increasing, and hence since 1970 the standard of living of blacks as a group with respect to that of whites has declined.

In 1980 blacks made up about 12 percent of the U.S. population. From 1971 through 1980 the median income for all black families, corrected for the effects of inflation, decreased by 8.3 percent. The median income for all white families also decreased over the same period, but only by 2.7 percent. Thus the long-standing economic disparity between blacks and whites in the U.S. increased. In 1971 the average income of all black families was 60 percent of that of white families, but by 1981 the figure had dropped to 56 percent.

The overall economic loss was not distributed equally in the black population. The real income of black families that included both a wife and a husband rose 7 percent in real terms over the 10-year period. The average income of such families in 1981 was about $20,000. The average income of all black families headed by a woman without a husband present, however, decreased by 8 percent. In 1981 the average income of such families was $7,500.

Much of the economic setback suffered by the black community as a whole in the 1970's was due to the fact that the proportion of families headed by women increased. In 1972, 32 percent of black families did not include a husband; by 1982 the fraction had increased to 41 percent. In 1982 only 12 percent of white families were headed by a woman with no husband in the home.

Several factors contributed to the rise in the proportion of black families without a husband. Black women who have children are less likely to have been married than white women with children. In 1982, 32 percent of all black women who maintained families had never been married, compared with 11 percent of white women.

In addition blacks are much more likely than whites to be divorced, although in the past decade the divorce rate in both groups has risen substantially. The divorce ratio, or the number of divorced people for each 1,000 married people living with a member of the opposite sex, was 265 per 1,000 for black women in 1982, a sharp increase from the ratio of 104 in 1970. Among white women the divorce ratio rose from 56 to 128 in the same period. As a result of the divorce rate and the tendency for women to remain unmarried even after they have had children, 34 percent of all the children who lived with one parent in 1982 were black, although only 15 percent of all children in the U.S. that year were black.

The familial and economic difficulties of blacks intensify each other, with the result that a substantially greater number live in poverty than was the case a decade ago. In 1970 eight million blacks had incomes below the poverty level as it was then defined by the Federal Government. In 1981 nine million blacks lived in poverty. In both years the figures corresponded to 34 percent of the total black population. In 1981, 11 percent of whites had incomes below the poverty level.

What is the cause of such disproportionate hardship? The answer is that the removal of legal barriers has not eliminated the economic obstacles that keep blacks confined to a limited sector of the economy and cause them to be laid off first in times of economic recession. In 1982, when the unemployment rate was the highest since World War II, the rate of unemployment among blacks was 19 percent, or about double the rate of 9 percent that prevailed among whites. The ratio of two to one between black unemployment and white unemployment was roughly the same as it was in 1972, when the overall levels of unemployment in the U.S. were much lower. Thus in generally worsening economic conditions the historical disadvantages of blacks persisted.

At the end of the 1970's the blacks who had jobs were still concentrated at the bottom of the occupational hierarchy. According to the results of the 1980 census, blacks in the experienced civilian labor force worked mainly in three broad job categories: (1) operators, fabricators and laborers, (2) technical, sales and administrative-support occupations and (3) service occupations. The three groups accounted for 74 percent of the black workers.

In some of the more detailed occupational classifications utilized in the census blacks are heavily overrepresented. For example, although blacks make up only about 10 percent of the civilian labor force, more than half of all private-household cleaners and servants were black, as were one-third of all garbage collectors and one-fourth of all postal clerks, nurse's aides, hospital orderlies and hospital attendants. Conversely, only 3 percent of all experienced lawyers, physicians and engineers were black, and only 4 percent of all managers and administrators.

Reynolds W. Farley of the University of Michigan Population Studies Center and Suzanne M. Bianchi of the Bureau of the Census have concluded on the basis of the information in the current report and other surveys that the black community in the U.S. is being split into two groups: an upwardly mobile group with middle-class aspirations and a permanent, poverty-ridden underclass. They argue that the social and financial distance between the best-off and worst-off blacks has increased considerably in recent decades: "In the 1940's almost all employed blacks worked as farm laborers, unskilled factory workers or domestic help. As blacks became better educated and moved into higher-status professional areas, the prestige gap between the top and bottom quarters of the occupational ladder widened."

Farley and Bianchi observe that for the worst-off blacks things are grimmer than the data on occupation and employment alone would suggest. One group of the most oppressed are not included in figures on unemployment: those who have become so discouraged about finding a job that they have given up looking for one. According to Farley and Bianchi, the fraction of discouraged job seekers "has increased among white men from 3 to 5 percent, but it has more than doubled among black men, from 5 to 12 percent, between 1970 and 1980."

They conclude that "two signs point to trouble for the future of black Americans and for society as a whole. One is the growing number of working-age


THE GIFT OF LOVE
SARONNO
AMARETTO
di
SARONNO®
ORIGINALE
A LIQUEUR PRODUCED BY
ILLVA SARONNO ITALY
56 PROOF
PRINTED IN ITALY
AMARETTO DI SARONNO
ORIGINALE·1525
TO SEND ANYWHERE IN THE U.S. CALL (800) 328-6148 VOID WHERE PROHIBITED
LIQUEUR 56 PROOF SOLE U.S. DISTRIBUTOR, FOREIGN VINTAGES, INC., NEW YORK, N.Y. © 1993 UNIQUE PRODUCT FROM ILLVA SARONNO, ITALY

men who do not hold or seek regular employment; the other is the growing economic gap between husband-wife families and families headed by women. In 1960 one-quarter of all blacks in poverty lived in female-headed households; today the figure is 60 percent. The children who grow up in these families may be leading candidates for the next generation of poor and chronically unemployed, signaling an even greater division among black Americans in the future."

Cruising for Trouble

The 464 ground-launched cruise missiles (GLCM's) scheduled to be deployed by the U.S. in Europe beginning next month are representative of a rapidly proliferating class of weapons that threaten to make the arms race even less controllable than it is today. In addition to the intermediate-range, nuclear-armed GLCM's intended for deployment in Britain, Italy, West Germany, Belgium and the Netherlands, the U.S. is currently installing air-launched cruise missiles (ALCM's), also armed with nuclear warheads, on B-52 bombers stationed in the U.S. and is vigorously pursuing the development of an assortment of sea-launched cruise missiles (SLCM's), both nuclear and non-nuclear, for service on submarines and surface warships. Within a few years, it is projected, the U.S. will have augmented its already formidable strategic nuclear forces with thousands of highly effective cruise missiles mounted on a variety of mobile launching platforms around the periphery of the U.S.S.R.

Estimates of the status of cruise-missile developments in the U.S.S.R. vary, but it is generally agreed that the Russians are substantially behind the U.S. in this particular area of advanced military technology. As in the case of multiple independently targetable reentry vehicles (MIRV's), however, there seems to be little reason to doubt that the Russians will eventually be able to match the American achievements and hence be in a position to present a comparably enhanced threat to targets in the U.S. Thus the net result of the advent of the cruise-missile era is likely to be a further reduction in the national security of both sides.

From the standpoint of arms control the main drawback of cruise missiles arises from precisely those qualities that make such weapons militarily attractive in the first place: they are smaller, cheaper, easier to hide and more versatile than long-range ballistic missiles, and hence they lend themselves to more flexible and deceptive basing modes. (Although cruise missiles are much slower than ballistic missiles, this disadvantage is thought to be more than offset by the superior accuracy of their continuous-guidance systems and by their ability to fly low over the terrain, evading radar tracking.) Because cruise missiles are hard to identify reliably by "national technical means" of surveillance (chiefly satellite photography), they are expected to present unprecedented difficulties to any nation attempting to verify another nation's compliance with an agreement limiting their number. Restrictions on cruise missiles, based on novel "counting rules" and other special verification aids, were incorporated in the SALT II Treaty and Protocol, but they remain in abeyance, since the treaty has not been ratified by the U.S. Senate. Given the extraordinarily strict standards of verifiability espoused by representatives of the Reagan Administration, the outlook does not appear to be bright for the inclusion of significant controls on cruise missiles in any agreement that might emerge from the current START negotiations in Geneva.

Meanwhile the U.S. is already preparing for the arrival of the next generation of cruise missiles. Early this year the Air Force announced that it would halt the current ALCM-B program in 1984, cutting off production at approximately 1,700 missiles rather than the 4,348 previously planned. The decision is expected to facilitate the timely procurement of the "follow on" weapon, the ACM, or advanced cruise missile, which will incorporate several improvements over its predecessors: (1) a more efficient engine, giving the ACM a much greater range than the ALCM-B; (2) a combination of microwave-absorbing materials and smoother design features (referred to collectively as "stealth" technology) that will enhance the missile's ability to elude Russian defense radar, and (3) a new passive guidance system, called Autonomous Terminal Homing (ATH), that will make the new missile not only more accurate but also less vulnerable to detection. According to one report, the advanced cruise missiles of the late 1980's may be capable of operating at supersonic speeds over intercontinental distances, giving them the potential of serving as strategic "counterforce" weapons against "hard" targets such as missile silos and command centers.

Appropriate Medicine

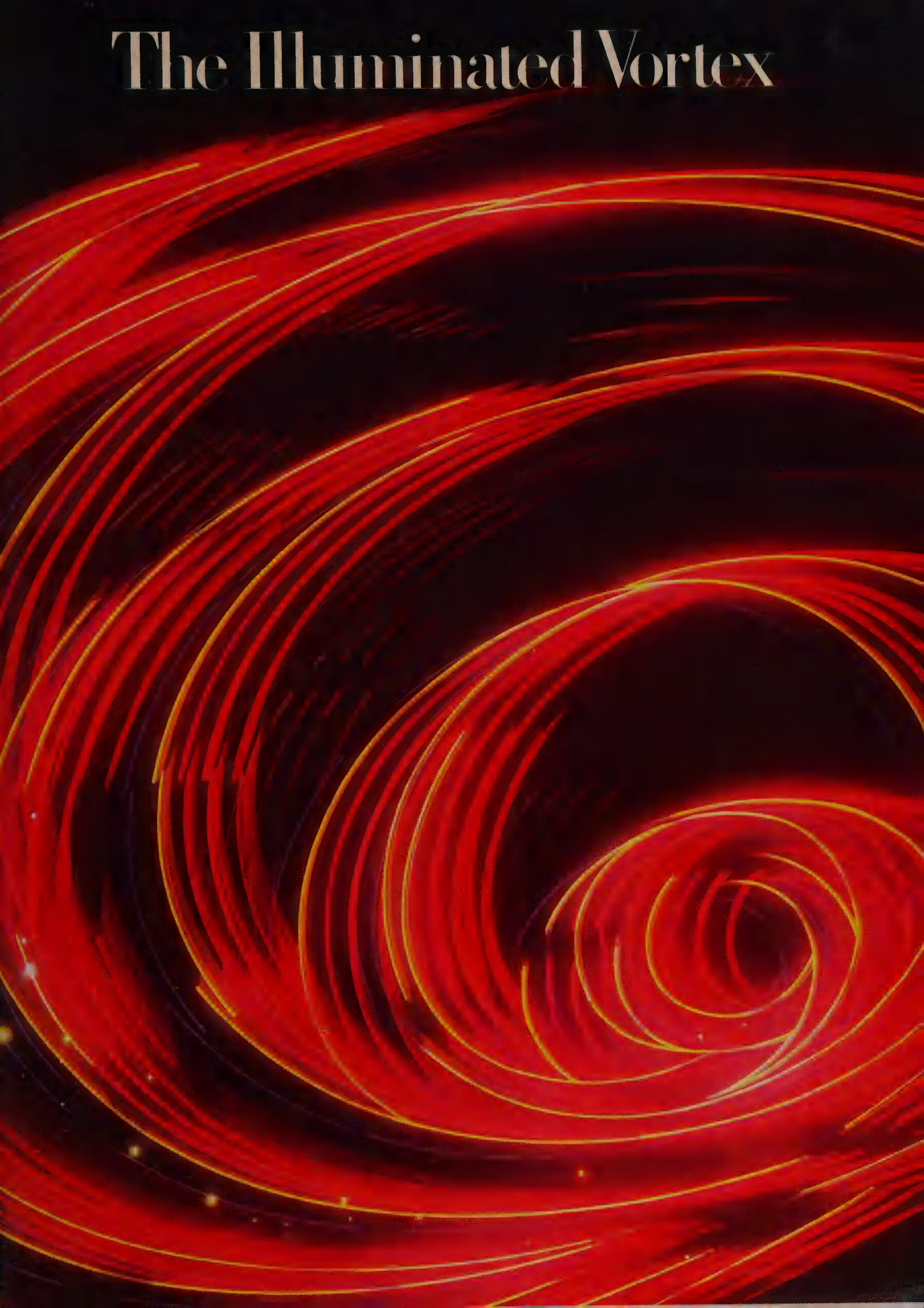
Diarrhea kills at least five million children a year, mostly in underdeveloped parts of the world, and contributes to the malnutrition of many millions more. It kills by dehydration. The treatment is obvious: replace the lost water and salts. This can be done by intravenous injection, but such therapy calls for sterile fluids, special equipment and skilled personnel, and it is expensive. It is generally unavailable precisely where childhood diarrhea is most prevalent. There is an effective, inexpensive alternative: oral rehydration therapy. A simple mixture of glucose and salts is mixed with local water and the solution is given by mouth; the patient's thirst regulates the intake to match what is lost in the watery stools. The potential of oral rehydration therapy has been clear for two decades, and recent data indicate that its routine application can virtually eliminate mortality from diarrheal dehydration. The task is to make the treatment available where it is needed.

The development and efficacy of the oral treatment are reviewed in an annotated bibliography published by the Pan American Health Organization and the World Health Organization. Acute diarrhea can be caused by a variety of agents, most of which are spread in contaminated water. Whatever the agent, the essential outcome is usually a reversal of the normal functioning of the small intestine. Sodium, and with it water, fails to be absorbed as usual from the gut into the bloodstream. Instead water and essential salts are secreted into the gut and are excreted in the stools. Even after it became clear that replacement of water and salts can save even most victims of cholera, the severest diarrheal disease, it seemed that only intravenous injection could circumvent the inability of the intestinal wall to absorb any orally administered fluid.

The key to oral therapy came more than 30 years ago. As early as 1949 it was found that glucose and some other simple sugars dramatically increase the absorption of sodium and water by the intestinal wall. A solution of glucose, sodium chloride (ordinary salt), potassium chloride and sodium bicarbonate in water could be given by mouth as maintenance therapy after initial rehydration by the intravenous route. In time it became clear that the oral treatment alone would suffice in most cases. The WHO developed a packet containing the correct amounts of glucose and the salts to be dissolved in a liter of water. The packets can be distributed through health-care systems and can be made available to families for administration in the home. If the packets are not available, it has been shown that a simple salt and sugar solution (a pinch of salt and a small handful of sucrose, or ordinary sugar) is often enough.

Oral rehydration therapy is a good example of effective "appropriate technology," in which laboratory investigation and clinical trials have developed a scientifically sound method suitable for widespread and inexpensive application under even primitive conditions. Commenting on an international conference on oral rehydration held in June in Washington, *The Lancet* recently pointed out that the problem is now one of inadequate implementation. Physicians, including those serving poor populations in developed countries, need to be

The Illuminated Vortex



The Illuminated Vortex

Understanding how the in-cylinder flow of the fuel-air mixture is influenced by chamber geometry provides a key to improving engine performance. By applying a laser measurement technique, a researcher at the General Motors Research Laboratories has gained new insight into the behavior of the flow.

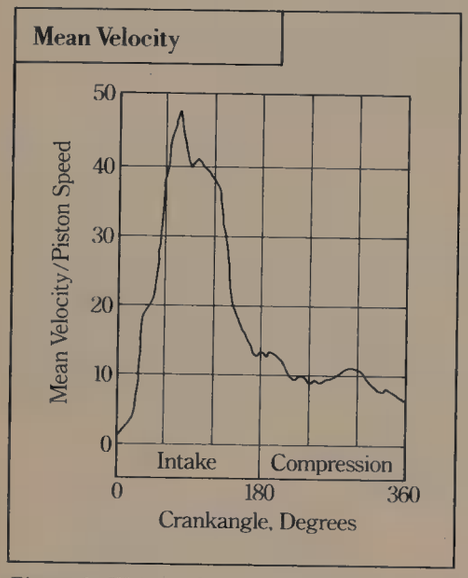


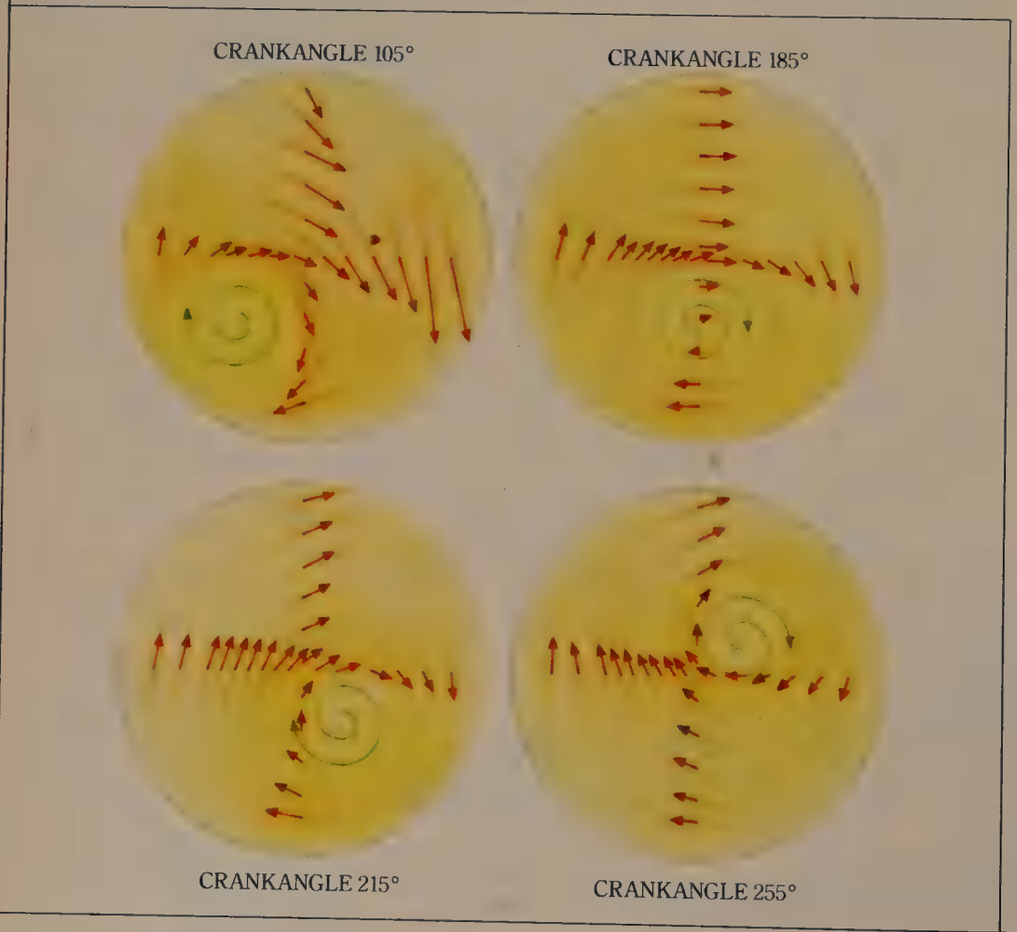
Figure 1: History of mean velocity at a single engine location.

Figure 2: Panoramic view of engine flow patterns. With changing crankangle, the center of rotation precesses from the cylinder's lower left quadrant to its upper right quadrant.

THE FLUID motions inside engine cylinders have considerable influence over the progress of combustion. Mixing of air and fuel, combustion rate, and heat losses from the cylinder are all important transport processes strongly dependent on fluid motions. The motion inside the cylinder has two components. Mean velocity influences the transport of momentum, energy, and species on a cylinder-wide scale, while the turbulence component influences the same phenomena on a local basis. The in-cylinder flow field depends primarily on the geometry of the cylinder and inlet port. Hence, decisions made in the engine design stage exert a controlling influence over the flow. But before questions about how different geometrical features affect the flow field can be answered, the problem of how to measure the flow must be solved. By applying Laser Doppler Anemometry (LDA), Dr. Rodney Rask, a researcher at the General Motors Research Laboratories, has obtained detailed measurements of the flow field.

LDA is a technique in which two focused laser beams pass into the cylinder through a quartz window. In the minute measuring region where the laser beams cross, a regular pattern of interference fringes is created. As the 1-micron particles, which have been added to the engine inlet flow, cross the measurement region, they scatter light in the bright fringes. In Dr. Rask's LDA system, the scattered light is collected by the same lenses used to focus the laser beam, and measured by a photomultiplier tube. The resulting signal is processed electronically to determine the time it takes a particle to traverse a fixed number of fringes. Since the fringe spacing is a known function of the laser beam crossing angle, this transit time provides a direct measure of velocity.

During operation of the LDA, measurements of velocity as a function of engine rotation (crankangle) are made at a number of locations within the cylinder. The instantaneous velocity at each point must then be separated into mean and turbulence components. The simplest technique is to declare that the mean velocities for all cycles are identical and ensemble average the data. However, this approach ignores the cyclic variation in the mean velocity. Another technique looks at individual cycles and uses a variety of methods, including sophisticated filtering, to split the instantaneous velocity into its components. This



approach is consistent with the LDA measurements, which clearly show that the mean velocity does not repeat exactly from one engine cycle to the next.

Differences in the flow field from one cycle to the next can seriously compromise engine efficiency. Near the end of the compression stroke, it is important to maintain a consistent velocity at key cylinder locations (e.g., at a spark plug). Dr. Rask's LDA measurements have identified design features that control cyclic variability.

FIGURE 1 shows mean velocity measured at a single location during an engine cycle. High velocity exists during the intake stroke when the inlet flow is rushing through the narrow valve opening. This jet-like flow into the cylinder causes large velocity differences between adjacent cylinder locations and produces strong turbulence. As the end of the intake stroke is approached (180 degrees in Figure 1), the levels of both mean velocity and turbulence drop rapidly. This decrease is a result of the changing boundary conditions for the cylinder—from strong inflow to no inflow. During the compression stroke the flow field evolves, but it undergoes no drastic changes. However, in a high-squish chamber, where the flow is forced into a small bowl in the piston or cylinder head, considerable turbulence is generated near the end of the compression stroke.

Measurements from many cylinder locations are necessary to make the flow field understandable. Figure 2 shows four flow patterns covering a period from near the end of intake into the compression stroke. Note the strong vortical flow, with the center of the vortex away from the cylinder center and precessing with changing crankangle.

By experimenting with geometrical variables, Dr. Rask has gained new understanding of phenomena observed in operating engines. The resulting knowledge has guided the design and development of new engines with a minimum of trial-and-error testing. The LDA findings are also being used to validate and calibrate engine flow computer models under development.

"From our measurements," Dr. Rask states, "we have been able to deduce how changes in the geometry of the port and combustion chamber modify the velocity field. These flow field effects are now being used to help designers tailor engine combustion for optimum performance."

General Motors

GM

MARK OF EXCELLENCE

THE MAN BEHIND THE WORK



Dr. Rodney Rask is a Senior Staff Research Engineer in the Fluid Mechanics Department at the General Motors Research Laboratories.

Dr. Rask received his undergraduate and graduate degrees in mechanical engineering from the University of Minnesota. His Ph.D. thesis concerned the Coanda effect.

Prior to joining General Motors in 1973, Dr. Rask worked on the design of nuclear reactors at the Knoll's Atomic Power Laboratories. In addition to further refinements in LDA measurement techniques, his current research interests include computer simulation of engine systems, with special emphasis on the intake manifold.



You can count on 3M diskettes. Day after day.

Just like the sun, you can rely on 3M diskettes every day. At 3M, reliability is built into every diskette. We've been in the computer media business for over 30 years. And we've never settled in. We're constantly improving and perfecting our product line, from computer tape and data cartridges to floppy disks.

3M diskettes are made at 3M. That way, we have complete control over the entire manufacturing process. And you can have complete confidence in the reliability of every 3M diskette you buy.

Look in the Yellow Pages under Computer Supplies and Parts for the 3M distributor nearest you. In Canada, write 3M Canada, Inc., London, Ontario. If it's worth remembering, it's worth 3M diskettes.



3M hears you...

3M

made more aware of the serious impact of childhood diarrhea and of the efficacy of oral treatment. Educational materials are needed to show mothers how to administer the treatment and to discourage reliance on traditional antidiarrheal medicines, most of which do more harm than good. The salient need is the development of effective networks for the delivery of primary care to provide the necessary education and materials.

Quarked Spark

What makes the sun shine? The short answer is that thermonuclear reactions deep in the sun's interior release energy in the course of fusing hydrogen into helium. A more complete account, however, must specify the pathways of the reactions, and the only direct evidence bearing on the pathways is the flux of neutrino by-products from the intermediate reaction steps. In principle such a flux can be detected on the earth because neutrinos interact with matter so rarely that they pass through the outer layers of the sun virtually unimpeded. Since 1967 Raymond Davis, Jr., of the Brookhaven National Laboratory has measured the neutrino flux at the bottom of the Homestake Gold Mine in Lead, S.D., and his results have presented a profound puzzle: the measured flux is lower than the predicted one by a factor of four. Now four physicists at Ohio State University have come forward with a new pathway for the nucleosynthesis of helium that may resolve the discrepancy. Richard N. Boyd, Ronald E. Turner, Michael C. T. Wiescher and Lawrence J. Rybarcyk invoke the presence of exotic atomic nuclei near the center of the sun whose electric charge, instead of being some integral multiple of the charge of the electron, is fractional. They describe their work in *Physical Review Letters.*

Fractionally charged particles must be extremely rare in nature if they exist at all; the idea that all electric charge is integral has been accepted as one of the basic properties of matter ever since Robert A. Millikan measured tiny electric charges early in this century. According to the theory of quantum chromodynamics (QCD), however, all nuclear matter is made up of quarks, which bear a charge of 1/3 or 2/3 times the charge of the electron. In most interpretations of QCD theory isolated quarks cannot exist; the quarks bind together to form particles with integral charge. The proton, for example, is made up of two "up" quarks, each with charge 2/3, and one "down" quark, with charge $-1/3$; the net charge is 1. The existence of bound states of quarks whose net charge is nonintegral, however, is not ruled out by QCD; indeed, such particles may have been found. In 1981 William M. Fairbank, George S. LaRue and James D. Phillips of Stanford University reported the detection of small concentrations of nonintegrally charged particles in the metal niobium. Although the Stanford findings have not been replicated, the sensitivity of the experiment is so high that the results cannot easily be dismissed.

The Ohio State theorists suggest that nonintegrally charged nuclei in the sun could catalyze a cycle of helium nucleosynthesis whose existence has not previously been recognized. Until now two major reaction pathways have been held responsible for the production of solar energy. The dominant pathway begins with the fusion of two protons to form a deuterium nucleus, and so the reaction is called the proton-proton reaction. It results in the emission of a neutrino, but the average energy of such neutrinos is too low to be detected by Davis' apparatus. Subsequent steps in the reaction, however, release neutrinos of much higher energy; it is primarily these neutrinos that can be observed. According to the standard theoretical model, reaction pathways beginning with the proton-proton reaction account for about 98 percent of the helium nucleosynthesis in the sun and for 98 percent of the sun's energy.

The second major reaction pathway is the carbon-nitrogen-oxygen (CNO) cycle. It was first proposed in 1939 by Hans A. Bethe of Cornell University and independently by C. F. von Weizsäcker in Germany. Nuclei of carbon 12 in the sun's interior combine with three protons, one at a time, to form increasingly heavy nuclei. The accretion of protons continues until it has built up a nucleus of nitrogen 15, consisting of seven protons and eight neutrons. (Two of the accreted protons decay into neutrons; each decay is accompanied by the emission of a barely detectable neutrino.) With the addition of a fourth proton the growing nucleus splits into a helium nucleus and a nucleus of carbon 12, whereupon the cycle begins again. In the standard model of solar burning the CNO cycle accounts for the remaining 2 percent of the sun's energy output.

The flux of high-energy neutrinos that can be measured on the earth varies widely for relatively small changes in the temperature at which the solar reactions take place. The Ohio State workers therefore realized that one way to reconcile the observed neutrino flux with the predicted one is to assume a lower temperature for the region of the sun where the neutrinos are emitted. If the assumed temperature is reduced by a million degrees Kelvin, from 14.4 to 13.4 million degrees, the production of neutrinos whose energies are high enough to be detected falls off by a factor of four. Without some additional reaction pathway for producing helium, however, such a temperature would also lower the

Play the finest

HARPSICHORDS & FORTEPIANOS

Build it yourself from one of our kits or let our experienced craftsmen build it for you.

Three Centuries of Harpsichord Making by Frank Hubbard (Harvard Univ. Press. 397 pgs. $20.00 postpaid)



For brochure send $1.00 to:
FRANK HUBBARD INCORPORATED
144-S Moody St., Waltham, MA 02154
(617) 894-3238

Published on the centennial of the eruption

Krakatau 1883

The Volcanic Eruption and Its Effects

TOM SIMKIN AND RICHARD S. FISKE

An indispensable reference including:

- **88 eyewitness descriptions of the phenomena**
- **28 reprinted scientific explanations and interpretations**
- **much of Verbeek's 1885 book (translated into English for the first time)**
- **27-page chronology of events**
- **474-entry bibliography and a name/subject index**

464 pp. 139 figs., 49 full-color illus.
Paper ISBN 841-0 $15.00

Order from:
Smithsonian Institution Press
P.O. Box 1579
Washington, D.C. 20013

Available October, 1983. International orders accepted.
All orders must be prepaid by check, money order, or credit card.

Oceanus®

The International Magazine of Marine Science and Policy

With offices in one of the world's leading marine research centers and published by the

Woods Hole
Oceanographic Institution,

Oceanus magazine is in a uniquely favorable position to monitor significant ocean research and related policy issues.



A case in point is the recent discoveries of deep ocean hydrothermal vents at various locations in the Pacific Ocean. These discoveries are causing a revolution in Oceanography.

Many long-standing biological and geological theories are being revised as a result of this exciting research. Exploration of the vents by the submersible Alvin is turning up new life forms (derived chemosynthetically rather than photosynthetically) and possible new sources of metals, such as zinc, iron, and copper. Scientists are getting a first-hand look at how the earth breathes while garnering supporting evidence for the theories of sea-floor spreading and plate tectonics.

Economically, it is a time of great challenge, too. The cost of running submersibles and support ships, for example, is high and rising. And federal funding for marine research all the while is shrinking. Many critical decisions must be made in nearly every area of oceanic concern.

Oceanus is meeting this challenge by publishing articles that help our readers grasp the significance of present research and that expose them to the substance of important public policy questions.



Recent thematic issues have dealt with:

- Deep Sea Mining
- Research Vessels
- Sharks
- Coastal Problems
- Pollution
- Oceanography from Space

Upcoming issues will look at:

- Marine Birds
- Oil & Gas Activities
- Industry and the Sea
- Polar Research

and much, much more.

Handsomely illustrated with many sharp photographs and line art, **Oceanus** is edited for all those with a serious interest in the three-quarters of our planet that is covered by water. The general readership includes many students, educators, representatives from government and industry, environmentalists, conservationists, as well as scientists in the disciplines of marine biology, geology and geophysics, chemistry, ocean engineering, and physical oceanography.

In addition to the thematic editorial material, **Oceanus** also features:

PROFILES of selected oceanographers that give a glimpse of the human side of science.

A CONCERNS SECTION that exposes the reader to controversial subjects and opinions.

A BOOK REVIEW SECTION that gives readers a convenient way to choose new and timely reading material.

LETTERS TO THE EDITOR or a forum for debate on views and thoughts expressed in the magazine.

Won't you join the Editors of **Oceanus** — published quarterly — in the exploration of the last great wilderness on earth, our oceans. Subscribe today!

Oceanus

SUBSCRIPTION ORDER FORM

Please enter my subscription for:

- ☐ one year at $20.00
- ☐ two years at $35.00
- ☐ payment enclosed ☐ bill me

(prepayment preferred)

Please send MY Subscription to:

Name (please print)

Street Address

City State Zip

Please make checks payable to Woods Hole Oceanographic Institution. Checks accompanying foreign orders must be made payable in U.S. currency and drawn on a U.S. bank. (Outside U.S., add $3 per year to domestic rates). Mail to: OCEANUS, Woods Hole Oceanographic Institution, Woods Hole, Massachusetts 02543.

HUGHES



INSURE OPTIMUM RESULTS

The people behind Hughes Support Systems are talented . . . determined . . . proud. They lead the challenge to new frontiers of technology, from a newly created, ultra-modern, working environment.

At Support Systems, Hughes people show the world how to use the systems of tomorrow . . . providing a full compliment of essential support services and equipment. Including training simulators, portable and depot test systems. Field and logistics support. Technical training for customers and development of training manuals. Covering programs ranging from electro-optics, lasers and missile systems, to airborne radar.

At Hughes, there is a stimulating relationship between the people and their work. A relationship that provides opportunity for substantial individual contribution. That's what Hughes Support Systems is all about. People.

People like you.

With Electro-Optics test experience (12 years + preferred):

Sr. Scientists — Responsible for application and development of EO test and measurement devices. Systems expertise in FLIRs, detectors, lasers, imaging devices, fiber optics, MS or PhD in EE or Physics required.

Project Managers — Management responsibility for technical performance, cost and scheduling in development of EO test sets. Plus direct feasibility and development-mode efforts. Project management background and BS in Physics or EE required. MBA desired.

Send your resume to:
Hughes Support Systems,
A1/1B404, Dept. 017,
P.O. Box 9399, Long Beach,
CA 90810-0463.

Creating a new world with electronics

HUGHES

HUGHES AIRCRAFT COMPANY

Proof of U.S. Citizenship Required.
Equal Opportunity Employer.

SUPPORT SYSTEMS

energy output of the sun by 22 percent. The proposed pathway restores the emitted solar energy to its observed value without releasing neutrinos whose energy is high enough to be detected.

The nonintegrally charged nuclei play the same role in the new reaction as carbon-12 nuclei play in the CNO reaction: they catalyze the accretion of free protons. The reaction begins with what is called a quarked helium nucleus, made up of two protons, two neutrons and an "up" quark; the nuclear charge is $2\frac{2}{3}$. Three protons are added to this nucleus one at a time until a particle made up of three protons, four neutrons and an "up" quark has been built up. Such a particle is essentially a lithium-7 nucleus combined with an "up" quark. With the addition of the fourth proton the quarked lithium-7 nucleus splits into a helium nucleus and a quarked helium nucleus, and the cycle begins again. Just as in the other two reactions, the essential change is the fusion of four protons into a helium nucleus, with the release of energy and neutrinos. According to the Ohio State group, the reaction can proceed in such a way, however, that the energy of the neutrinos is too low for them to be detected on the earth.

The newly proposed reaction pathway can generate the missing 22 percent of the sun's energy if nonintegrally charged nuclei make up only one part in 10^{15} of the sun's interior. The proportion is consistent with the error limits given so far by the Stanford experimenters. Nevertheless, the model must be viewed with caution. The Stanford result is not fully accepted, and quarked nucleosynthesis can account for the missing neutrinos only if physical quantities such as the internal temperature of the sun fall within extremely narrow limits. According to William A. Fowler of the California Institute of Technology, the proof of any theory of solar-energy production must await the construction of a more sensitive neutrino detector. One such detector that has been proposed would essentially consist of 50 tons of the rare metal gallium. The main obstacle to its realization is that this amount of the metal is more than the world's annual production and would cost $25 million.

Catalytic Cup

The shape of many proteins is central to their function. Enzymes, for example, offer binding sites with particular shapes so that particular molecules are brought into positions favoring a particular chemical reaction. The reaction would otherwise be improbable. The question arises: Can a chemist design a molecule that has a useful shape, so that it facilitates a reaction as an enzyme does? In an article in *Accounts of Chemical Research* C. David Gutsche of Washington University in St. Louis describes his effort to synthesize molecules that resemble the bell-shaped Greek vase known as the calix crater. He calls them calixarenes.

By a quirk of chemistry the methods Gutsche employs are reminiscent of a synthesis developed by Leo Baekeland at the beginning of this century. Baekeland heated formaldehyde (HCHO) and phenol (C_6H_5OH), a six-carbon ring in which five of the carbon atoms are linked to hydrogen atoms and the sixth is linked to a hydroxyl group (OH). The result was a hard, resinous material: the first commercial plastic, Bakelite. Gutsche and his colleagues start with a phenol that has a second substitution: the hydrogen atom across the ring from the hydroxyl group is replaced by a tertiary butyl group (C_4H_9), which consists of three methyl groups (CH_3) bound to a central carbon atom. When this compound is treated with formaldehyde, it forms calixarenes, a mixture of cyclic substances: mostly four, five, six or eight phenol rings forming larger, cup-shaped molecules.

Each phenol ring is one wall of such a cup. The butyl groups protrude above



its rim; the hydroxyl groups jut inward at its base, closing off the cup at its bottom. In essence, then, each calixarene circumscribes a central cavity. The cavity acts as a trap; thus the calixarene has a "propensity to form molecular complexes with smaller molecules." The trapping depends on the size of the cavity. The four-sided calixarene *p-tert*-butylcalix[4]arene forms complexes with chloroform, toluene and pyridine, the five-sided calixarene *p-tert*-butylcalix[5]arene forms complexes with isopropyl alcohol and acetone, the six-sided calixarene *p-tert*-butylcalix[6]arene forms complexes with chloroform and methyl alcohol. X-ray crystallography confirms that the small molecule is indeed sequestered inside the central cavity. The calixarenes therefore simulate enzymes that bind a small molecule inside a pocket-shaped site.

How well might a calixarene mimic an enzyme? Gutsche and his colleagues are considering the enzyme aldolase. In animals it serves an early stage in glycolysis, the metabolism of glucose. For the most part it catalyzes the splitting of the six-carbon molecule fructose 1,6-diphosphate into two three-carbon molecules, dihydroxyacetone phosphate and glyceraldehyde 3-phosphate. In green plants it does the reverse, thus serving a late stage in photosynthesis. Its action is known in detail. First an amino group (NH_2) protruding outward from the amino acid lysine, a part of the enzyme, binds the middle carbon of dihydroxyacetone phosphate. The binding is assisted by the electrostatic attraction between the negative charge of the phosphate group of the three-carbon molecule and a positive charge on the enzyme. Next the three-carbon molecule is stripped of a proton (a hydrogen nucleus). Its resulting negative charge facilitates a linkage to glyceraldehyde 3-phosphate, which aldolase also has ways to capture.

Gutsche and his colleagues hope to graft three features of aldolase onto a four-sided calixarene some of whose sides have been induced to flip downward so that the calix has what they call a partial-cone conformation. An amino group is to be stationed partway up the exposed inner surface of the cup; it is intended to mimic the ability of aldolase to bind the middle carbon of dihydroxyacetone phosphate. An ionized metal atom is to be stationed at the bottom of the cup; it is intended to attract the phosphate group of dihydroxyacetone phosphate. A basic group (CO_2^-) is to be stationed at the rim of the cup; it is intended to strip a proton from dihydroxyacetone phosphate. The simplest ways to synthesize calixarenes cannot produce such features, and so Gutsche and his colleagues are devising more complex synthetic techniques.

The Wonders of Sian

Ever since a well digger exposed the first of some 6,000 life-size terra-cotta soldiers buried two kilometers from a great earth tumulus in southeastern Shensi in the summer of 1975, it has been apparent that the Chinese had struck an incomparable archaeological bonanza. The tumulus was traditionally held to be the tomb of China's first emperor: Ch'in Shih Huang Ti. His burial place, prepared years before his death in 210 B.C., had been the subject of speculation, much of it the stuff of legend, ever after. Until 1975, however, digging in the enormous mound by would-be thieves had revealed nothing, and no official excavation had been attempted.

Since that time archaeologists from the Chinese Academy of Social Sciences, which is responsible for historical research, and from the Shensi provincial government have been excavating elsewhere near the great tumulus, adding to the list of recovered grave offerings hundreds of life-size terra-cotta warriors and horses, together with bronze weapons and other artifacts. One of the most recent finds consists of the first



Wear the Expanding Universe on your chest.

Photo courtesy of Hale Obs.

Let the world know about your interest in astronomy with T-shirts and bumper stickers from the ASTRONOMICAL SOCIETY OF THE PACIFIC.





ASTRONOMERS DO IT AT NIGHT

SUPERNOVAE ARE A BLAST

Since 1889, the A.S.P. has brought together professional astronomers, amateurs, educators and interested laypeople. Although we began as a Pacific Coast organization, today our members can be found in every state and 54 foreign countries.

Membership in our non-profit society is open to everyone with an interest in astronomy. Benefits include: *Mercury* magazine, which brings you authoritative non-technical articles on the latest discoveries about the Universe; a directory of astronomy activities and services in your area; and a membership card good for discounts at planetariums and stores around the country.

The A.S.P. also makes available a wide variety of astronomical materials, including photographs taken with large telescopes and NASA spacecraft, as well as many books, tapes, and charts about the heavens. For example, the Society has prints and slides of the beautiful images from the flyby's of Jupiter and Saturn.

For a free illustrated brochure and a complete catalog, please send the coupon (*below*) to:

A.S.P., Dept. S
1290 24th Ave.
San Francisco, CA 94122

Please send your free brochure and catalog:

Name: _____

Address: _____

_____

Zip

bronze effigies found so far: two teams of four chariot horses, with their harness virtually complete, each team drawing a two-wheeled bronze chariot complete with a driver, arms outstretched to hold the reins, clad in a long-sleeved winter cloak and wearing a cap covering his ears and neck and decorated with a cockade. The effigies are half life-size; one of the two has recently been placed on display at the museum erected on the site not long after the first discoveries.

Ch'in Shih Huang Ti holds a unique place in Chinese history. According to tradition, his royal house was instructed in about 770 B.C. to police the western Chou capital city of Anyang after the Chou ruler, then the most powerful of China's feudal kings, had moved his capital to the east. Over the next 150 years the Ch'in state grew to be the most powerful in western China, and by 338 B.C. it was also the best-governed in all China. Shih Huang Ti succeeded to the Ch'in throne in 246 B.C., some 15 years after his predecessor had effectively destroyed the military power of the paramount feudal state, the Chao. (Chou kings, although they were still given some lip service, were by then powerless.) Young Shih Huang Ti also went campaigning, and between 246 and 221 B.C. he brought every other feudal state in China to its knees. With his power consolidated, he declared himself emperor of all China and went on to effectively dismantle the feudal systems of his former rivals and establish his own central authority. He partitioned the nation into 40 provinces, which he ruled through his own civil service: inspectors who oversaw the activities of local magistrates.

Fearful of death, he prepared heroically for his afterlife. Legend has it that within the mountainous tomb he had raised a few score kilometers from his capital (modern Sian) was a great relief map of his empire, with mercury endlessly flowing along its river channels. When the emperor died at the age of 49 in 210 B.C., legend further relates, his burial chamber was filled with molten copper so that his coffin was sealed within a single huge metal ingot.

Chinese officials state that the bronze chariot-team effigies, found 20 meters outside the western edge of the great tumulus, are the most important find in the vicinity of the tomb since the initial discovery of the terra-cotta army in 1975. Considering that some 12 square kilometers, including the tumulus itself, remains to be explored, the Shensi site represents one of the great archaeological enterprises of all time. The excavation of Herculaneum and Pompeii, the towns buried by an eruption of Vesuvius in A.D. 79, has been in progress since 1748; the clearing of Shih Huang Ti's immense funerary monument may well take longer.



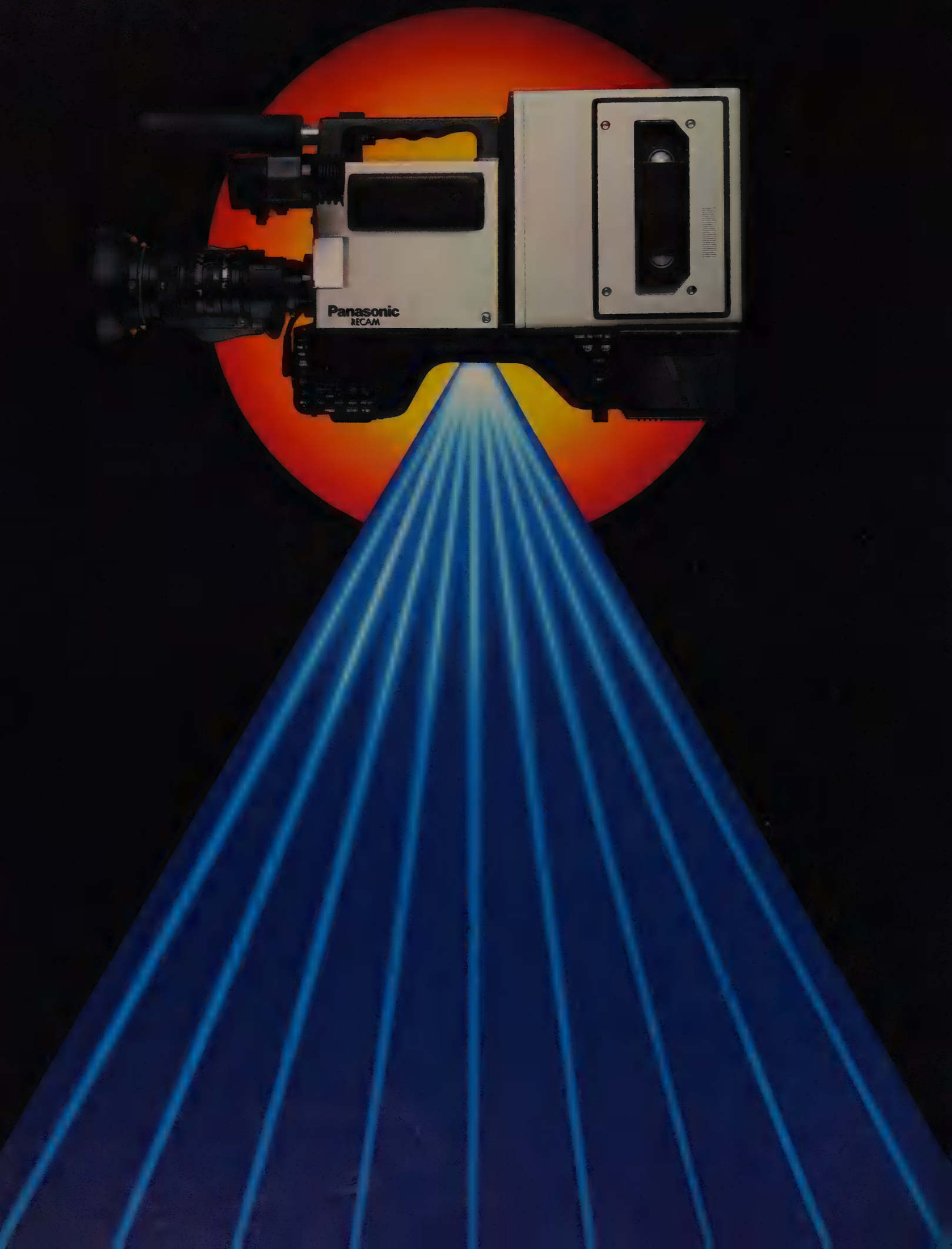
JAPANESE TECHNOLOGY TODAY

In Japan, the presentation itself is as important as what is being presented. Here, the formal Kimono and traditional Meishi-Bon symbolize the importance of the Laser Semiconductor Diode, one of the new technologies helping to revolutionize information storage and retrieval. Such technological developments are ushering in the New Industrial Revolution.

SCIENTIFIC AMERICAN

Advertisement

MATSUSHITA IN VIDEO TECHNOLOGY:



How to build video cameras that can go to the top of Mt. Everest, the bottom of Lake Ontario, or inside Hollywood's next epic.

Video. The catchword of the '80's. Taking our eyes places they may never have seen before. Suddenly, it's possible to gaze from the top of Mt. Everest, to scan the bottom of Lake Ontario, to examine a nocturnal world that the human eye only records as pitch black. And Matsushita (pronounced Mot-SOOSH-ta) Electric, Japan's largest consumer electronics company, is making video cameras that show us the light.

Close your eyes for one second. And take a step.

Then transport yourself to 29,028 feet above sea level and take another. Where the sky meets the jet stream. Where winds whip your face at 84 miles per hour. Where your oxygen intake is only two-thirds of what it was before. Your legs feel like lead. Your brain is disoriented. Your eyes cannot focus.

This is Everest. This is where on May 7, 1983, Panasonic video cameras made history.

VIDEO THAT MET THE JET STREAM.

ABC Sports wanted to open their 19th season of "The American Sportsman" with first-time videotape coverage of an expedition reaching the summit of Mt. Everest. They demanded video cameras with uncompromising picture quality. Everest demanded absolute portability. Panasonic gave them both.

With Recam,™ one of the Panasonic cameras on the climb, video technology has reached the state of the art. The recorder is built right into the camera saving up to half the weight of earlier components. Now a cameraman can shoot without being tied to a technician carrying a separate tape deck. The result: Unheard-of mobility. And videotape footage never captured before.

As explained by Julius Barnathan, President of Broadcast Operations and Engineering at ABC, "Recam offers great flexibility in difficult shooting situations."

With Recam, the sky is literally the limit. And that's where ABC took it. On a hot-air balloon expedition over the jungles of Tanzania, giving them mobility to shoot locations no video camera crew could easily reach.

But a primary use for Recam could be prime time, itself. Already, WNEV-TV Boston, has become the first major network station to go directly on the air using this compact ½-inch tape. The cost savings are enormous. And that rates highly with any TV station.

Recam can play a variety of roles, but one of the most notable will be in Hollywood. Where Universal Pictures is using Recam to shoot a behind-the-scenes look at the making of "Dune." A science-fiction epic that's targeted to be one of the biggest productions in Hollywood history.

VIDEO THAT WENT BACK INTO TIME.

While Recam is reaching heights on Mt. Everest and in Hollywood, Matsushita has another camera that's reaching new lows. 300 feet below the surface of Lake Ontario. Where National Geographic Society scientists recently used a Matsushita color video camera as the "eyes" in an expedition to help locate and record two American gunships sunk during the War of 1812. The depth and icy waters made for extremely difficult diving conditions. But this remote-controlled camera was able to locate both gunships in near pristine condition.

VIDEO THAT'S AHEAD OF ITS TIME.

While some of our video cameras are operating in the deep, other Matsushita cameras are operating in the dark. Designed for industrial security, the Matsushita Stellar Lite® cameras use our patented Extended-Red Newvicon® tubes allowing them to see in as little as 25/1,000,000ths of a footcandle. Pitch black to the human eye. They're being used in areas too expensive or impractical to light. Like airport aprons and sports stadiums.

But Matsushita is more than video. It's over 14,000 different products sold in more than 130 countries under the *Panasonic, Technics, Quasar* and *National* brand names. Consolidated worldwide sales in 1982 were over $14 billion. And in the world of business, as in the world of video, that's a beautiful picture.

MATSUSHITA ELECTRIC
PANASONIC TECHNICS QUASAR NATIONAL

JAPANESE TECHNOLOGY TODAY



THE AUTHORS

The 1983 edition of "Japanese Technology Today" was written by Dr. James C. Abegglen and Mr. Akio Etori.

Dr. Abegglen is Vice-President and Director of the Boston Consulting Group, Inc., and concurrently Professor of International Management at Sophia University, Tokyo. He has written numerous articles and books including *The Japanese Factory,* a pioneering study of Japanese corporate organization.

Mr. Akio Etori is an award-winning and respected writer on science and technology, having covered these areas for over 20 years. He is managing editor of *Saiensu,* the Japanese language edition of *Scientific American* and Japan's leading and most prestigious publication in this field.

THE COVER

Photographer Arthur Beck captures the importance of one of the latest developments in component technology from Matsushita Electric. Presented on the traditional Meishi-Bon, the new Terraced-Substrate Laser Semiconductor has made possible such information processing devices as optical memory disk systems, digital audio disk systems, and laser printers in computer terminals.

Art direction and supervision of text and illustration were provided by Ted Bates Advertising/New York.

The dynamism of the Japanese economy over the past 30 years, and the impact of that dynamism on the world, must be considered a pivotal phenomenon of the 20th century. For the first time since the industrial revolution, a non-Western nation has moved to the highest levels of output and income. The world now has a new, and unexpected, source of capital and technology that bids to overshadow the traditional Western sources. An English observer has suggested that just as the historical period 1775–1875 might be termed the British Century, and 1875–1975 the American Century, so the next hundred years may come to be seen as the Pacific Century. If so, the driving force will have been Japan.

The magnitude of Japan's change is very great indeed, and in a short period of time. As recently as 1960, only 20-odd years ago, the Japanese economy made up less than 3 percent of the economy of the world. A 3 percent economy may be treated with condescension by a 33 percent economy, as indeed Japan was treated by the United States in those days. *(See Table 1.)* A 3 percent economy may provide occasional nuisances in trade relations, but is hardly likely to be the object of systematic protectionism. Nor is such an economy likely to be the subject of admiring studies of its management systems.

In 20 years, the Japanese economy became a 10 percent economy. (Some license is needed here to accommodate current aberrations in exchange rates, which now somewhat understate Japan's position.) A 10 percent economy attracts very serious attention. This is more the case when its principal trading partner has over the same period experienced a very substantial reduction in relative economic position. While Japan moved from insignificance to major importance economically, the United States' share of world GNP moved from about one-third to about one-fifth.

This kind of change in both absolute and relative position is very hard to accommodate. The issues raised in terms of economic relations, especially trade, are clear enough. More subtle issues of changes of attitudes on the part of the Japanese, and toward the Japanese, are, nonetheless, real problems. The difficulties of adjustment are seen in trade tensions, in problems of redefining strategic relations, in efforts to analyze and better understand the Japanese accomplishment and its sources, and in Japan's own problems of internal social and political adjustment to the consequences of drastic economic change.

Even in the midst of these tensions and readjustments, there is still in Japan a considerable propensity to look ahead, to gauge probable directions and problems, as well as to attempt to measure outcomes. The current view is that Japan will continue to outperform the other developed economies to the year 2000. *(See Table 2.)* That is, the relative economic position of the U.S. is seen as continuing to decline, but at a slowing rate, with Japan continuing to grow at a relatively rapid, but again slowing, rate. In this process per capita output in Japan will come to be greater than that in the U.S. by a significant margin.

These continuing changes in the size of the total economy have implicit in them very great changes in the structure of the economy. Such structural changes have been taking place, and at a rapid rate, as indicated in *Table 3.* There has been a massive shift of the Japanese labor force out of the primary sector of agriculture, forestry and fisheries. The shift was initially into the construction and manufacturing sector through the 1960's and early 1970's. It is already clear, however, that a further shift is in process in recent years, with employment in manufacturing declining relative to employment in the tertiary sector of distribution, utilities, services and the like. That is, the structure of the Japanese economy, in terms of employment, is already moving in the direction of services that has been conspicuously the case with the U.S. in recent years.

As Japanese analysts look forward, they see this shift in the structure of the economy accelerating. As is indicated in *Table 4*, the level of employment in the primary sector is expected to continue to decline rapidly. By the year 2000 less than 5 percent of the Japanese labor force will be in farming, forestry and fishing. (With this, there is very likely to be a steady reduction in political resistance in Japan to greater imports competitive with these primary industries.)

Of even greater interest in terms of the pattern of expected growth is the view that there will be very little increase in employment in the manufacturing indus-

TABLE 1

SHARE OF WORLD GNP
1960–1980

	1960	1970	1980
US	33.7%	30.2%	21.5%
EEC	17.5	19.3	22.4
Japan	2.9	6.0	9.0
USSR	15.2	15.9	11.6
PRC	4.4	4.9	4.7
Other	26.3	23.7	30.8
Total	100.0	100.0	100.0

Source: 2000 Nen No Nihon (Japan in the Year 2000), Keizai Kikaku Cho Sogo Keikaku Kyoku Hen, Nihon Keizai Shimbunsha, Tokyo, 1982, p. 31.



"The Honda Accord is far and away the best compact in the land. Its engine hums contentedly. Its ride is smooth and quiet. Its interior is a model of good taste and expensive-looking accoutrements." (*Car and Driver,* May 1982)

"In short, the Honda Accord gets my vote for best all-around car under $10,000." (*Washingtonian,* January 1983)

"From the standpoint of product value, though, Honda has given the new model what it takes to capture the attention of thoughtful buyers." (*Road & Track,* January 1982)

"It is to Honda's everlasting credit that the cars are such all-around sweethearts to drive." (*Car and Driver,* January 1982)

"They are beautifully put together, very comfortable, and rewarding to spend time with." (*Car and Driver,* January 1982)

"The new 4-speed automatic overdrive transmission merits high marks." (*Motor Trend,* June 1983)

"Hard cornering is something the Accord does quite well for a family sedan. After all, why should driving cease to be pleasureful just because you have a family?" (*Motor Trend,* January 1982)

"The Accord is one of the best sedans of any size that money can buy." (*Car and Driver,* May 1982)

"An acknowledged benchmark." (*Motor Trend,* June 1983)

"How do you follow a benchmark?" (*Car and Driver,* January 1982)

HONDA

The Accord 4-Door Sedan

© 1983 American Honda Motor Co., Inc.

JAPANESE TECHNOLOGY TODAY

tries in Japan between now and the end of the century. Indeed, employment in most sectors of manufacturing is expected to decline, with the machinery sector alone showing substantial increase. This sector, of course, takes in the whole range of electronics and other high-technologies.

The dramatic change in employment patterns over the period to the end of the century is expected to be in services. In fact, this sector is expected to account for essentially all of the increase in employment over the period. The employed population is expected to increase by about eight million and employment in services is expected to increase by a like amount. Services is a broad category, of course. It includes auto repairs and beauty parlors, architects and advertising agencies, computer advisers and tax consultants. The significance of this projection lies in the view that Japan's economy will be shifting over the next two decades very rapidly from an emphasis on hardware to a focus on software, to the highest value added sector of the economy.

The underlying causes of this expected shift in employment patterns can be seen in *Table 5*, showing expected changes in the composition of the economy. Such sectors as chemicals and metals will make up only a small part of the Japanese economy at the turn of the century. Machinery manufacturing will largely take the place of these declining sectors. At the same time, services will move from about one-third of GNP to nearly one-half of GNP.

It is against this broad economic background that the present situation and future prospects of Japanese technology must be seen. Technology is both cause and result of these structural changes in the economy. The structural changes themselves are the cause of rapid growth. Technology makes the changes possible. In turn, the changes demand new technologies. In this cycle of change, increases in productivity are possible, and the shift to higher value added is accomplished.

The successive oil shocks and their consequences have imposed on all economies an urgent requirement for structural change. The transitions are difficult for all economies. Japan has managed the transitions with more speed and thoroughness than most. Further, Japanese analysts, as we have seen, expect a series of continuing structural changes in the economy—indeed, welcome them despite the problems they present—since it is with these changes that income levels are raised, and standards of living improved.

Japan has by definition in this post World War II period of unprecedented economic growth managed a number of restructurings of industry. The first was in coal mining, which in its peak years in the early 1950's employed more than 450,000 workers in more than 900 mines and has since the early 1970's employed fewer than 50,000 workers in less than 50 mines. In this same period Japanese coal imports went from virtually nil to some 80 percent of total consumption.

Following mining, the next very large industry that came under pressure as the Japanese economy advanced was the textile industry. Here the effect has been more that of increased productivity than decreased output as employment in fibers and fabrics dropped from nearly 1.5 million in the mid-1960's to less than 900,000 by the end of the 1970's. Imports of fibers, fabrics and apparel have increased sharply, the more so since Japan now has the lowest tariffs on textiles in the world, and maintains no Multi-Fiber Agreement import restrictions.

TABLE 3

THE STRUCTURE OF JAPAN'S LABOR FORCE (Percent)

	Agriculture, Forestry, and Fisheries	Mining, Manufacturing, Construction	Utilities, Transport, Services
1955	37.6%	24.4%	38.1%
1960	30.2	28.0	41.8
1965	23.5	31.9	44.6
1970	17.4	35.2	47.4
1975	12.7	35.2	52.0
1980	10.4	34.8	54.8

Source: Japanese Ministry of Finance

A more recent and dramatic case of industry restructuring in Japan was that of shipbuilding, where with the collapse of world demand from 1974, the industry in Japan, with government support, moved to cut capacity. Japan had 50 percent of world shipbuilding capacity, at the peak of demand. In two years in the late 1970's, 40 percent of that capacity—or 20 percent of world capacity—was closed down.

For all of these past successes the task facing Japanese industry and government in the 1980's to achieve the restructuring necessary to meet national aspirations is a formidable one. The context needs to be sketched briefly. By the late 1960's it began to become apparent to many in Japan that the industries on which Japan had built its initial successes as rapid growth began in the mid-1950's were in the process of becoming inappropriate as Japan sought higher levels of productivity and value added. These industries included many that had earlier been singled out for rapid development, such as petrochemicals and non-ferrous metals. They had formed the core of Japan's shift from labor intensive sectors like textiles into more sophisticated capital-intensive industries.

Still, in the emerging context of the late 1960's and early 1970's these raw-material-intensive, pollution-intensive, land-extensive industries were of less efficiency and relevance. This was the more true as the raw material supplying countries, Japan's trading partners, sought to put more value added to the materials before exporting them—smelting bauxite to alumina or to ingot before exporting, for example. It was in Japan's interest to shift these industries offshore, both in terms of better using Japan's own resources of capital and labor, and in terms of raising the level of trade with supplying countries.

The abrupt and catastrophic increase in

TABLE 2

WORLD GNP, YEAR 2000 (Percent)

	1980		2000		Annual Growth
Advanced Economies	63%		58%		2.8%
Japan		10%		12%	4.0
US		22		20 }	2.5
EEC and Other OECD		31		26 }	
Developing Economies	15		20		4.6
NICs		4		7	6.0
LDCs		11		13	4.0
Socialist Economies	22		22		3.2
USSR		13		12 }	3.0
Eastern Europe		5		5 }	
PRC		5		5	4.0
World Total	100%		100%		3.2

Source: Japan in the Year 2000

See beyond the ordinary.

The Canon A-1 is no ordinary camera. It is a creative tool. Conceived as the ultimate in automatic SLR's, the A-1 is unsurpassed in providing exposure control options. There are six, to be precise, allowing you to select the one best suited to your subject. Choose a shutter speed to control and interpret action. Select a lens opening and blur away a background.

In the programmed mode, the A-1 makes both of these decisions for you so you can really concentrate on your subject. You just focus, compose and shoot.





And create. A bright digital display in the viewfinder shows the speed and aperture in use, whether the choice was yours or the camera's.

The A-1 provides the versatility to match your imagination. Add any of over fifty Canon FD lenses. A Canon Speedlite for automatic flash. You can shoot at up to five frames-per-second with the optional Motor Drive MA. But most important, the A-1 does everything automatically. Freeing you to shoot a special subject in your own special way, and make a picture that nobody else saw.

The Canon A-1. It's half of what you need to turn photography into fine art.



The Official 35mm Camera of the 1984 Olympic Games

Canon® A-1

Canon USA, Inc., One Canon Plaza, Lake Success, New York 11042
140 Industrial Drive, Elmhurst, Illinois 60126 / 6380 Peachtree Industrial Blvd., Norcross, Georgia 30071
123 Paularino Avenue East, Costa Mesa, California 92626 / 2035 Royal Lane, Suite 290, Dallas, Texas 75229
Bldg. B-2, 1050 Ala Moana Blvd., Honolulu, Hawaii 96814 / Canon Canada, Inc., Ontario

© 1983 Canon USA, Inc.

energy prices in the 1970's turned what had been seen as a long-term process into one of great urgency. For among their other characteristics these industries are energy-intensive and notably unsuitable for siting in Japan, with the world's highest energy costs and a notably unreliable energy supply. Thus in the early 1980's Japan confronts a massive problem of industry restructuring of the greatest urgency. Something of the nature of the problem can be appreciated from the data in *Table 6*. A series of important industries, employing upward of 250,000 persons directly, with several hundred corporations, are in deep and abiding difficulty.

There is general agreement in Japan that the scale of these industries must be reduced, and resources that might be devoted to maintaining these industries will be better deployed elsewhere. Indeed, it might be argued that the Japanese are especially persuaded followers of the Schumpeterian concept of "creative destruction." Perhaps because of the dynamism of their economy, where economic cause and effect can be observed directly and in a short period, there seems to be in Japan an unusual appreciation of the fact that movement to higher standards of living does in fact require that some industries are losers in the process while others become the next generation of winners.

As an aside, it might be noted here that these cases are useful in providing some perspective on and perhaps a partial antidote to the currently popular literature in the U.S. that uncritically extols Japanese management methods. All of the companies in all of these industries are cases of "Japanese-style management." The oftencited characteristics of Japanese management—consensual decision-making, concern with quality, special attention to employee needs and the like—do not help these industries survive. Or put more precisely, whatever the considerable merits of many aspects of Japanese practices, these management methods are not able to override fundamental economics of a company's or industry's position. Further, when discussing the Japanese economy and its corporations, these cases should serve as a helpful reminder that Japan too has winners and losers, that no amount of management skill, government competence, or labor quality can prevent, or should be allowed to prevent, uneconomic companies and industries from confronting basic economics.

TABLE 4

STRUCTURE OF EMPLOYMENT
(Thousand Persons)

	1980		2000		Annual Change
Primary Sector	5,770		3,080		-3.1%
Secondary Sector	19,250		21,110		0.5
Manufacturing		13,770		14,200	0.2
Chemicals		1,750		1,450	-0.9
Primary Metals		670		540	-1.1
Machinery		5,380		8,930	2.6
Other Manufacturing		5,970		3,280	-3.0
Construction		5,480		6,900	1.2
Tertiary Sector	30,190		39,120		1.3
Utilities		300		330	0.4
Finance, Insurance, Real Estate		1,910		2,410	1.2
Transport and Communications		3,500		3,550	0.1
Services, etc.		24,480		32,830	1.5
Total	55,360		63,290		

Note: Population aged 15 to 64 years: 1980, 67,390; 2000, 67,230.
Source: Keizai Shingikai

It is this process of continual restructuring that poses Japan's technology challenge. Just as the movement out of the "cheap" trinkets and textiles of Japan's earlier, poorer era required the technologies that established petrochemicals and metals industries, so now newer technologies must be put in place to continue the move to yet higher levels of value of output. As we have seen, the generally held view of Japan's future assumes that these transitions will be accomplished and that the technology is in place already or will be forthcoming. It is, however, a formidable challenge.

The challenge is the greater since the process is a continuing one. It is now beginning to impact an industry in which Japan has for some time held uncontested technological and economic leadership—the steel industry. The quite exceptionally high level of Japanese steel-making technology can be seen from the data in *Table 7*. Japanese steel companies are now suppliers of technology to nearly every U.S. steel-making concern. The U.S. government's import controls—the trigger price mechanism—have assumed that Japan

TABLE 5

COMPOSITION OF GROSS DOMESTIC PRODUCT
(Percent)

	1980		2000	
Primary Sector	3.7%		4.2%	
Secondary Sector	38.2		31.6	
Manufacturing		29.3		21.6
Chemicals		5.4		1.5
Primary Metals		3.6		0.8
Machinery		11.9		15.7
Other		8.4		3.6
Construction		8.9		10.0
Tertiary Sector	58.1		64.2	
Utilities		3.0		1.5
Finance, Insurance, Real Estate		15.5		8.5
Transport and Communications		6.6		5.6
Services, etc.		33.0		48.6
Total	100.0		100.0	

Source: Keizai Shingikai; Choki Tembo Iinkai

MAXELL ADVANCES STATE OF THE ART TECHNOLOGY TO A HIGHER STATE.

WITH THREE CLEARLY SUPERIOR PRODUCTS.

Over the years, Maxell has been the leader in the audio/video/computer media business. And the advanced tape technology used to create our three new products will keep us on top for years to come.

For the new XL-S recording tape, we developed a superfine Epitaxial formulation. This greatly increases its dynamic range, making the XL-S the ideal tape for all high performance recording needs and perfect for digital audio disc recording.

We also developed a totally new Epitaxial formulation for HGX Gold video tape. Its superfine particles deliver brighter chroma levels and higher audio sensitivity. And the binding process, using molecular fusion, gives our new VHS and Beta videocassettes unmatched durability.

Gold Standard floppy disks are widely accepted as the choice for critical new disk drive testing. But the new 5¼″ high density disk, equal to an 8″ double sided, double density disk, with a capacity in excess of 1.6 megabytes, took our involvement one step further. Developed in conjunction with a new drive design, it helped redefine 5¼″ capacity.

Behind the technical excellence of Maxell audio, video and computer media is the human element—Maxell people. We're dedicated to creating the kind of superior product performance that brings out the best potentials of high end equipment. In fact, we insist on it. Shouldn't you?

maxell.
IT'S WORTH IT.



© 1983 Maxell Corporation of America,
60 Oxford Drive, Moonachie, N.J. 07074

TABLE 6

CONDITIONS IN PRINCIPAL DEPRESSED INDUSTRIES, 1981

Industry (Number of Companies)	Employees (Number)	Capacity Utilization (Percent)	Operating Profits (¥Billion)
Petrochemical (12)	137,013	59.0%	–57.7
Chemical Fertilizer (34)	1,200	62.0	–30.5
PVC (17)	1,887	54.8	–47.5
Electric Furnace Steel (58)	31,300	80.0	–11.5
Steel Alloys (31)	6,978	51.0	–1.6
Aluminum Refining (5)	5,844	59.0	–56.3
Paper and Pulp (600)	67,500	79.8	27.9
Synthetic Fibers (18)	29,928	82.7	0.7

Source: Asahi Shimbun 9 December 1982.

was the world's most efficient steel-producing country.

Yet Japanese steel exports peaked several years ago. There have been no new capacity additions, and none are planned. According to current press reports, the Japanese steel industry is considering dumping and countervailing duty suits against foreign steel suppliers. Imports began a few years ago, and in 1982 increased 41 percent over 1981 to the considerable total of 19 million metric tons. In the first seven months of 1983, imports were up another 25 percent over the same period in the preceding year.

Whatever the validity of the industry's complaints about dumping, and they leave curious echoes in Western ears, the fact is that steel too will have its day in Japan, and will then yield pride of place to yet higher-growth, higher-technology industries. This is not to suggest that the industry in Japan will die; rather, it will be forced to focus on the more sophisticated steel products as South Korea, Brazil, Taiwan and other newly industrializing countries become in turn the lower-cost producers of standard grades of steel. For all of its notable success in management, the steel industry of Japan cannot escape the consequences of Japan's continued economic success.

No doubt in their turn too such current Japanese successes such as the automobile industry will encounter the problem of "creative destruction" that now looms over Japan's very successful steel industry.

The technology that has powered this continuing industrial restructuring has had its phases, just as the industry's changes have had. For industry, the shift broadly has been from labor-intensive to capital- and energy-intensive to technology-intensive sectors, as we will see below with Japan's electronics industry. In technology the transition has been from massive introduction of imported technologies, to adaptation and further development of those technologies, to the present situation where Japan must in greatly increased measure initiate its own technologies.

Japan's long dependence on the importation of technologies is shown clearly in its continuing deficit on its technology trade balance. Japan still pays out more than $3 for technology for every $1 it receives in payment for the export of technologies. *(See Table 8.)* This deficit is steadily diminishing over recent years as Japan has moved into surplus on new technology trade. That is to say, Japan's large current deficit in technology trade is the result largely of major contracts entered into some time ago. Its current account is favorable, one measure of progress in Japan's independent development of technology.

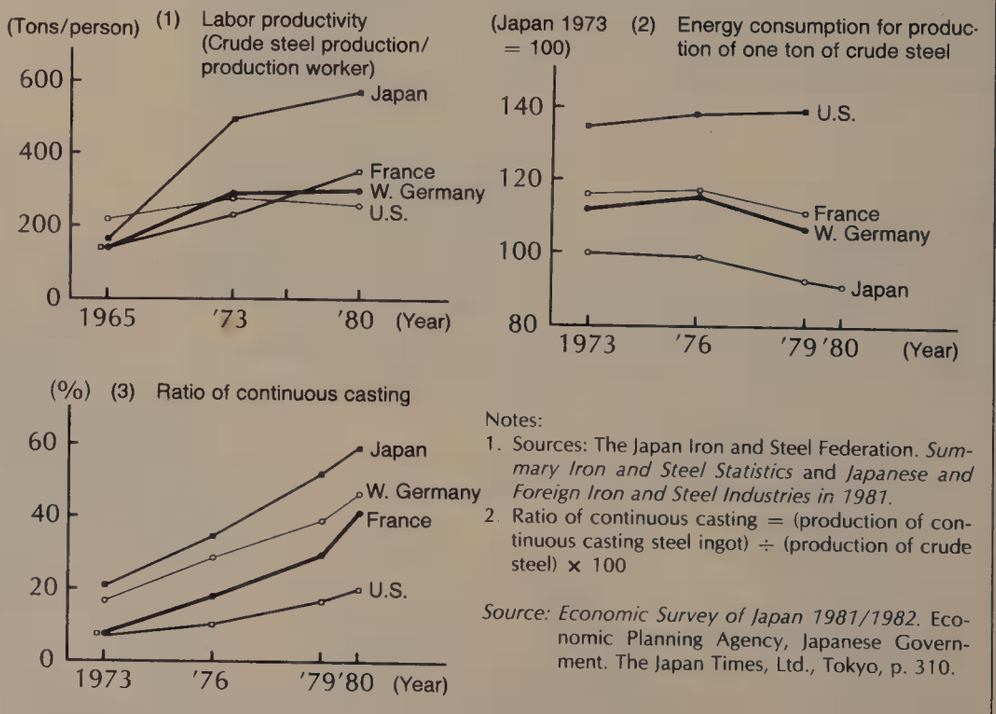
It is curious too that while the U.S., France and the U.K. have run surpluses on technology trade over the past decade, both West Germany and Japan have been in consistent and substantial deficit. No doubt some part of this is a function of catching up after wartime lags and destruction. Some part too, it has been suggested by the Japanese, may be the result of greater flexibility in these two countries and a greater willingness, and ability, to absorb technologies from abroad.

Japan's earlier dependence on imported technologies, and current need to move forward in independent development is very widely recognized in Japan both in public and in private circles. In the most recent *Science and Technology White Paper,* the issue was stated as follows:

"Our country has achieved very rapid progress since World War II, and in the advance of scientific knowledge as a technological revolution has contributed greatly to the progress of our economy and society. Further, in such sectors as electronics and industrial robots, a world-leading position has been attained. However, technology in our country as a whole depends largely on imported technology.

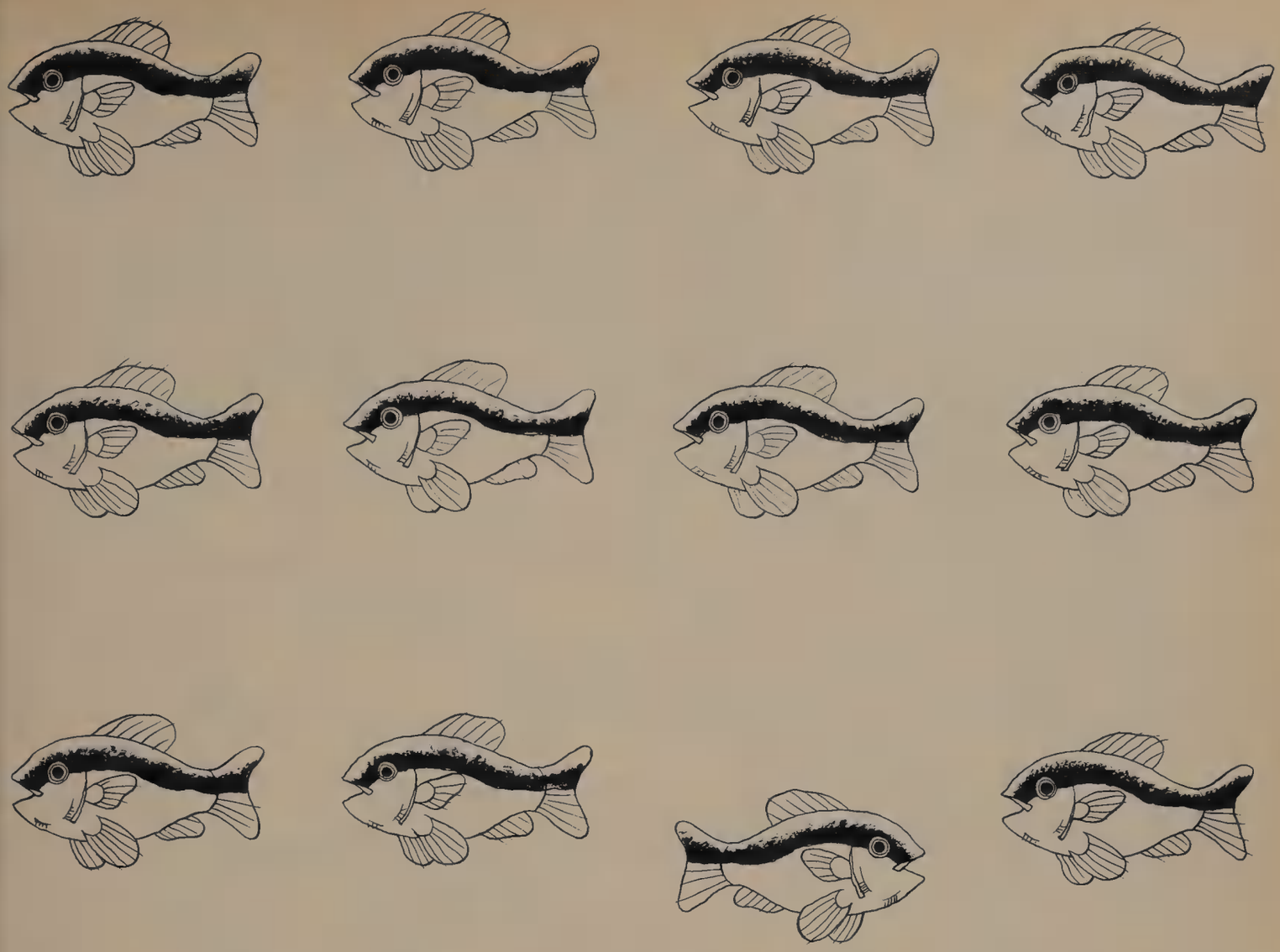
TABLE 7

A COMPARISON OF STEEL-MAKING INDUSTRIES OF SELECTED COUNTRIES



Notes:
1. Sources: The Japan Iron and Steel Federation. *Summary Iron and Steel Statistics* and *Japanese and Foreign Iron and Steel Industries in 1981.*
2. Ratio of continuous casting = (production of continuous casting steel ingot) ÷ (production of crude steel) × 100

Source: Economic Survey of Japan 1981/1982. Economic Planning Agency, Japanese Government. The Japan Times, Ltd., Tokyo, p. 310.



A leader should occasionally strike out in a new direction.

Through the years, Sumitomo
has often attempted the unconventional,
a strategy not always successful.
But the number of times that it
has been has helped Sumitomo grow
into the world's 6th largest steelmaker,
a company well respected
in the industry as a technological
innovator and developer of high-grade
specialty steels,
many unavailable elsewhere.
Today, Sumitomo Metals is going
beyond steel production, exploring
challenging new areas like
energy and electronics.
But that shouldn't surprise you.
We've been striking out
in new directions for years.

SUMITOMO METALS

SUMITOMO METAL INDUSTRIES, LTD. Tokyo & Osaka, Japan
Sumitomo Metal America, Inc. Head Office: 420 Lexington Ave., New York, N.Y. 10017, U.S.A. Phone: (212)949-4760 Offices in Los Angeles, Chicago, Houston.
MAIN PRODUCTS: Tubes & pipe, flat rolled products, bars, shapes & wire rods, rolling stock parts, steel castings & forgings, and a wide range of steel products.

Instead of owning a car maybe you need a cam



In a world full of cookie-cutter SLRs, Nikon introduces the world's

Over the years millions of people have bought millions of SLRs designed for beginners.

But what happens when you don't want to take pictures like a beginner anymore?

Many stick with their mass-produced SLR and wonder why they can't get better results.

For the few, however, who are determined to improve as photographers, Nikon has created an engineering marvel called the FE2. An automatic SLR so different, there are no other cameras to compare it with.

For instance, take the matter of Selective Focusing. Which is the ability to vary apertures to express your impression of the subject. An absolute must for anyone who wants to improve as a photographer.

The FE2 gives you more options for Selective Focusing than any other 35mm SLR in the world.

To begin with, the FE2 is an aperture-priority automatic. So you have total control over depth of field when you shoot in automatic.

But, as we said, that's just the beginning.

The FE2 is the first 35mm automatic with a 1/4000 shutter speed. An innovation made possible by Nikon's development of an ultra-thin honeycomb titanium shutter. It means photographers like yourself will, for the first time, be able to use a full range of apertures in bright sunlight. And because the FE2 is faster than any other 35mm automatic, it gives you greater control over exposure than other cameras when using the

If you love photography, don't miss Nikon's "The Photographer's Eye" on SPN Cable TV. Check local listings. For further information write Dept. 50, Nikon Inc., 623 Stewart Ave., Garden City, N.Y. 11530. © Nikon Inc. 1983.

era designed for millions
ra engineered for a few.



fastest 35mm SLR automatic. The FE2.

latest 1000 ASA color print film.

The FE2 is also the only 35mm automatic with a 1/250 flash synch speed. Which means you'll have the ability to select a wider range of lens apertures when taking flash pictures outdoors. An achievement that was once considered impossible with a 35mm SLR.

In addition, the FE2 is one of the few cameras equipped with through-the-lens flash metering (TTL). Which gives you unparalleled accuracy because flash exposures are measured by the camera as they pass through the lens. And not by a sensor on the flash unit itself. TTL allows you to achieve virtually unlimited aperture settings when shooting flash indoors with a dedicated Nikon Speedlight.

Even when you're not taking photographs with the FE2 it's an astonishing camera. Its balance and 'feel' make other SLRs seem bulky and unresponsive. And its copper silumin aluminum body makes it one of the most rugged Nikons ever made.

So while other cameras are designed for millions, the FE2 is engineered to take the kind of pictures only a few strive for.

One-in-a-million.

Nikon

We take the world's greatest pictures.™

In the flash of one second, Delta-10 can print the alphabet six times.

At 160 cps, 100% duty cycle, Delta-10 is built for speed. It's the dot matrix printer that quickly strikes up an intense working relationship with your computer.

There's no nonsense. Delta-10 performs with serial and parallel interfaces as standard. The 8K buffer gives Delta-10 the power to instantaneously store and print data as it frees your computer to continue its job. Turn Delta on and it can print three crisp copies simultaneously at the continuous speed of 160 cps.

And Delta-10's performance isn't only measured in swiftness. There's the diversity of its character fonts, its true descender matrix, and its underlining ability.

Star's new Delta-10. Designed for red-hot efficiency. And isn't that what you need?

star micronics · inc

THE POWER BEHIND THE PRINTED WORD.

Computer Peripherals Division
P.O. Box 612186, Dallas/Ft. Worth Airport, TX 75261



TABLE 8

TRENDS IN TECHNOLOGY BALANCE OF PAYMENTS
(Ratio: income/payments)

	United States	France	United Kingdom	West Germany	Japan Total	Japan Contracts
1971	10.56	0.85	1.07	0.39	0.12	0.71
1973	8.38	1.13	1.17	0.40	0.12	1.27
1975	9.09	1.27	1.02	0.42	0.23	1.42
1977	10.57	1.43	1.22	0.41	0.23	2.15
1979	8.77	—	1.29	0.46	0.27	1.94
1981	8.03	—	—	—	0.31	2.68

Source: Kagaku Gijutsu Hakusho (Showa 57 Nenhan) (Science and Technology White Paper). Kagaku Gijutsu Cho. Tokyo, 1983.

It might be said that developing on this basis has been a special feature of our economy.

"Our country has established a position as one of the advanced countries even in terms of technology. In order to deal with numerous problems both domestic and foreign, our country must free itself from dependence on imported technology, must open new paths to independent power regarding technology, and must press with greatest efforts creative research and development for finding and nurturing innovative technologies."

Given that Japan's social and economic ambitions require steady technological progress, and given that these increasingly depend on Japan's own efforts in development, a first issue is the scale of current investment in research. Two measures of Japan's investments in research are shown in *Table 9*: research expenditure and research personnel compared with other major economies. The U.S. spends by a great margin more on research than other countries. Japan's expenditure is about equal to that of West Germany, rather more than Britain and France combined.

However, Japan's rate of increase in research expenditure is much the highest of these countries, with West Germany next in rate of expenditure growth. An interesting, if somewhat academic, exercise is to note that if Japan and the U.S. continue to increase research budgets at the rate that they have over the past 15 years, by 1990 Japan's total research expenditure will exceed that of the U.S. Given straight-line projection of historical growth rates, the expenditure in 1990 for Japan would total ¥23,099 billion, and for the U.S. ¥21,325 billion. This would be a per capita expenditure for Japan twice that of the U.S. (The projection is fanciful since it assumes constant exchange, inflation and growth in expenditure rates.)

In terms of research personnel, Japan has about half the research personnel of the U.S. This suggests that on a per capita basis the labor effort in Japan in research is at about the same level as that in the U.S. It appears that Japan has about as many researchers deployed as do West Germany, the U.K. and France combined. By these measures, the Japanese research effort is now considerable, compares favorably with the research investments of the other major economies and is growing.

There is, of course, continuing and often heated discussion of the Japanese capacity to move to discovery and invention, as contrasted with application technology and adaptation. In a recent review of Japan's high technology industries, the Long-Term Credit Bank of Japan offered a balanced assessment:

"Several years ago it was a subject of debate among entrepreneurs, engineers, economists and other specialists as to whether the Japanese would ever take the lead in the newest technologies as they had done with conventional technologies.

"At that time the pessimists maintained a negative outlook for Japanese industry relative to frontier technology, for the following reasons. First there are no suitable grounds for invention and cultivation of original technologies in Japan. The inventive powers of the Japanese educational system and the R&D systems of private enterprises must be radically improved. In reality, however, such radical reforms are extremely hard to achieve. Second, Japan has no large-scale government-sponsored agencies like the U.S.'s NASA that support private industries' R&D efforts.

"Those pessimistic views were challenged by the positivists, however, who had ample evidence to the contrary. They argued that first, the Japanese have already invested in R&D efforts on the high technology frontiers and Japanese industries are energetic enough to meet the technological challenge of the advanced Western nations. In the past Japanese industry did not get enough government support for its basic research efforts, but recent years have witnessed the creation

TABLE 9

RESEARCH EXPENDITURE IN MAJOR COUNTRIES
(¥ billion)

	Japan	United States	USSR	United Kingdom	France	West Germany
1965	426	7,246	2,772 (1966)	911	730	716
1970	1,195	9,339	4,655	929	982	1,454
1975	2,622	10,438	6,885	1,410	1,813	2,293
1980	4,684	13,864	7,319 (1978)	1,463	2,241	4,899
CAGR*	17.3%	4.4%	8.4%	3.2%	7.8%	13.7%

RESEARCH PERSONNEL IN MAJOR COUNTRIES
(Thousand)

1965	117.6	494.5	664.6 (1967)	56.6	44.0	61.6
1970	172.0	540.9	927.7	54.7	57.3	78.4
1975	255.2	533.1	1,223.4	78.8	62.0	94.1
1980	302.6	643.5	1,373.3 (1978)	104.4	72.9	122.0
CAGR*	6.5%	1.8%	6.8%	4.2%	3.4%	4.7%

*Compounded Annual Growth Rate
Source: Professor K. Uno, Tsukuba University in *Nihon Keizai Shimbun,* 18 June 1983, p. 10.

of various government programs, such as the Creative Scientific Technology Promotion and Next-Generation Industrial Infrastructure Development. The outcome of these basic research programs will lay the foundation for their active application in the private sector, in the future. . . .

"As noted above, there are two conflicting views on the Japanese high-technology industry. However, judging from the recent spectacular growth of the semiconductor, robot, computer, optoelectronics and other high-technology industries, it appears that the Japanese will eventually gain an international edge in those industries. Still, there are various future technology areas, such as aerospace development, nuclear fusion, Josephson-junction technology, and discovery and invention of new principles where Japan will have difficulty catching up with advanced Western nations."

This assessment seems balanced and realistic. It should be added, however, that it provides an impression of more government support for research in Japan than the expenditure data indicate is warranted. The share of government in total research expenditures in Japan is notably low. The share of government expenditure in research is more nearly comparable, at least between Japan and the U.S., if defense-related research budgets are omitted from the figures. *(See Table 10.)* However, the U.S. defense budget does include substantial support for the semiconductor, computer, special metals and other industries' research. In fact, given U.S. budget pressures, the defense budget is no bad place to put a research subsidy. In contrast, given Japanese political attitudes and pressures, the defense budget is likely to be the last place in which to seek to get approval of research funding.

In any event, the U.S. nongovernment expenditure on research and development is, according to these sources, the equivalent of about ¥7.2 billion, while the nongovernment research expenditure in Japan is about ¥3.7 billion. On a per capita basis, with a population half that of the U.S., Japan's private sector R&D expenditure is now equal to that of the U.S.

These data provide only some measure of the effort that Japan has under way in developing technologies that will allow the process of continual renewal and restructuring of industries to take place. In fact, in the Japanese view progress in new technologies has been so rapid as to justify use of the term "new industrial revolution," in Japanese, *shinsangyo kakumei.* The term is widely used in Japan. It denotes in general new approaches to organization and production systems, first in the factory and now in the office as well. The core technology for this new industrial revolution, as the Japanese see it, is microelectronics and all of its applications through to what is being called the "information industry," the collecting, transmitting, analyzing and presenting of data.

Japan's electronics industry is rapidly on the way to displacing such traditionally critical industries as textiles and steel in the economy. *(See Table 11).* The long decline of textiles is clear and expected. The beginnings of the relative decline of the steel industry can be seen. The auto industry continues to increase its share of Japan's total manufacturing, but the fast growth is that of the electronics industry, now overtaking steel in the position it occupies in total manufacturing.

The industry is about equally divided in terms of sales value into three major categories: consumer electronics, industrial electronic products and electronic components. In a real sense the three sectors are mutually dependent, since developments and cost improvements in one sector can favorably impact other sectors. Further, as new approaches are developed, the interactions among the sectors become more evident. The computerized home will use the television set and the telephone for data transmittal. The components supplied to the consumer sector—integrated circuits and cathode-ray tubes, for example—gain volume and reduced cost for application to industrial products. It is an interactive industry, despite its disparate product applications.

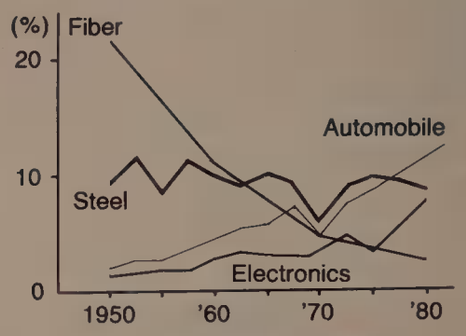
Japan's first successes in electronics were in the consumer sector. It is a sector of particular interest in terms of technology because it not only illustrates the steady restructuring of an industry to higher value added goods but also illustrates especially well the changing position of Japanese technology in terms of world competition.

Table 12 displays the shifts in Japanese consumer-electronics output over some three decades. In the immediate postwar period the industry was largely made up of the production of radios. Radio receiver production has for some years now largely moved out of Japan into the newly industrializing economies of East Asia, with Japan retaining production of only the most sophisticated sets. There are following a series of product life cycles, as monochrome television grew, reached market saturation, gave way to color television and declined. Again monochrome television has for some years now been produced largely in Taiwan and South Korea. Color television production as a percent of the total of consumer electronics in Japan has also had its day, and audio equipment went into a sharp decline in the past few years (to be revived by the new digital player technology?).

The consumer-electronics sector in Japan is now becoming dominated by videotape recorders, now making up nearly 40 percent of total consumer-electronics

TABLE 11

TRENDS IN COMPOSITION RATIO OF MAJOR MANUFACTURING INDUSTRIES



* Based on the shipment amounts in the Table on Manufacturing Statistics by the Ministry of International Trade and Industry, Japanese Government.
* The figures for the electronics industry exclude those of electronic desk calculators.

Source: Annual Data on Japan's Electronics Industry, 1983 Edition, Electronics Association of Japan, Tokyo, 1983.

TABLE 10

GOVERNMENTS' SHARE OF RESEARCH BUDGETS

	Total Research Budget (¥ trillion)	Government Share Including Defense Research Expenditure (percent)	Government Share Excluding Defense Research Expenditure (percent)
US	13.9 (1980)	48%	33%
West Germany	4.3 (1979)	44	41
France	2.2 (1979)	58	47
UK	1.5 (1978)	48	32
Japan	5.2 (1980)	28	28

Source: Journal of Japanese Trade & Industry, No. 4, 1983



Sony creates seventh row, center. Forever.

INTRODUCING THE SONY COMPACT DISC PLAYER.

You are looking at an invention so extraordinary that High Fidelity magazine has hailed it as "the most fundamental change in audio technology in more than eighty years."

A piece of audio equipment that permits you to hear something you've never heard before: perfection.

Sony is pleased to raise the curtain on the world's first digital audio compact disc player. The CDP-101.

THE END OF HI-INFIDELITY.

The CDP-101 uses a laser beam to play compact discs that hold far more musical information than conventional records ever could. The result: fidelity and depth of sound that no conventional turntable could ever deliver.

This highest of fidelity remains faithful too. Because the digital discs are read by laser beam, there is absolutely no physical wear. Nor is there any distortion from such annoyances as dust, fingerprints or even scratches. The laser simply sees through them.

Equally ingenious, an infrared remote control lets you select tracks without budging from your armchair.

While an ever-expanding library of compact discs lets you listen to your favorite artists as though you, and your armchair, were centered in the spotlight above.

Maybe most important, the CDP-101 is fully compatible with the conventional equipment you may own.

We suggest you hear the Sony Compact Disc Player soon. For a sound you can't believe, from the audio innovator you assuredly can.

SONY.

THE ONE AND ONLY



© 1983 Sony Corp. of America. Sony is a registered trademark of the Sony Corp. 1 Sony Drive, Park Ridge, New Jersey 07656.

output. The product has had an astonishing history. Production began in 1975. The compound annual growth rate of production in seven years since has been 96 percent, barely failing to meet the level of an annual doubling. Exports began in 1976, and have in fact somewhat more than doubled each year since, with a compound annual growth rate of 106 percent. The growth rate was at a more than doubling pace between 1980 and 1981 but slowed in 1982, no doubt due to depression-induced slowing of demand and necessary inventory adjustments. News reports suggest demand is surging again in 1983, both in Japan and abroad.

The product progression is an interesting one in terms of technology. Clearly consumer electronics has moved very rapidly to higher technologies across these several decades. Clearly too Japan was dependent on foreign sources for virtually all of the technology employed even to the stage of color television. RCA licenses made Japanese color-television production possible. Corning Glass supplied glass-tube technology to Japan's Asahi Glass, and lived to regret the fact bitterly and loudly. The early semiconductor production in Japan for the whole range of consumer electronics was available to Japan under license from Western Electric, Fairchild, Texas Instruments and others who held the basic patents. Perhaps more than any other of Japan's industries, the development of Japan's consumer electronics industry is the result of imported technology in the hands of aggressive and highly competitive Japanese firms that adapted, improved, drove costs down and came to dominate world markets.

It is hardly necessary to note as well that Japan's consumer electronics industry has been the subject of especially bitter trade complaints with recurrent charges of dumping, copying, conspiracy, government subsidy and the like. Quotas have been imposed by the U.S. and other countries. Japanese producers have found it necessary to set up production in Western Europe and the U.S. to avoid such trade barriers (which explains in part the life cycle configurations in *Table 12).*

It is this background that makes the case of the videotape recorder the more interesting. This is essentially Japanese technology. Indeed, Japan very nearly monopolizes world production, with its share of world VTR output generally estimated at more than 90 percent. Only Philips and associated Western European companies are in VTR production, and they have contrived to impose import quotas on Japanese VTR exports to the EEC market. There is no VTR production by U.S. producers. Instead, U.S. consumer electronics firms source VTR from Japanese producers for sale under their own brands in the U.S. market. The VTR thus represents a full turn of the wheel, a Japan-developed product with trade friction minimal because Japan is virtually the sole producer.

Several themes implicit in this case have disturbing implications for U.S. and European trade positions: the unwillingness of U.S. producers to risk investment in this new product, the speed with which the Japanese moved in developing the VTR market and investing accordingly, the dominant position this success threatens to provide Japanese producers as VTR technology is extended to cameras and films and other applications. In only 30 years Japanese consumer electronics has moved from marginal production of the cheapest and lowest-technology radio receivers to a dominant position in a massive new product area.

If consumer electronics as an industry segment is especially visible, and therefore especially trade-sensitive, so electronic components may be considered the basis for an overall position in the industry. It has been suggested frequently that the semiconductor industry is the equivalent in the late 20th century of the steel industry in an earlier period of industrialization—the basic industry that must be in place in order to maintain a strong economy. Certainly the explosion in microelectronics is based on semiconductors, and advances in semiconductor technology largely determine the direction and rate of developments across the whole range of electronic-related products.

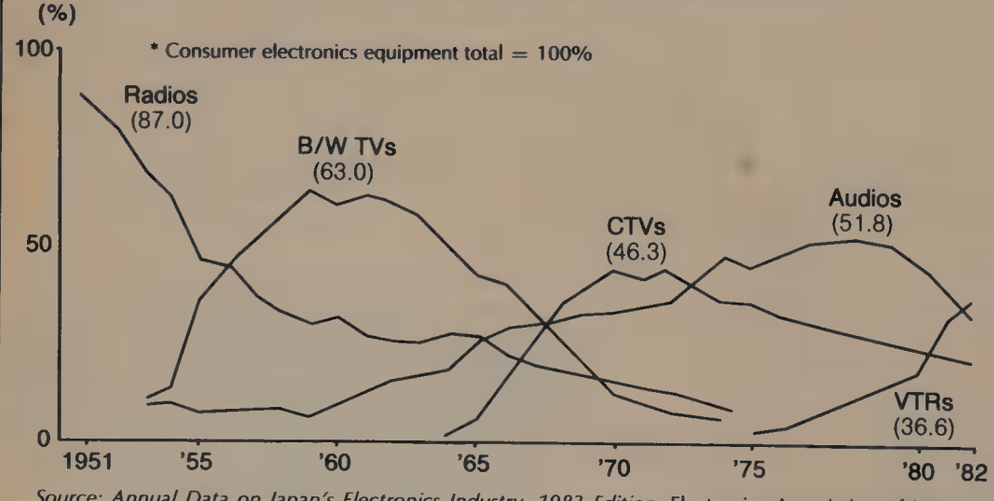
Table 13 reviews the Japanese experience to date in integrated circuits. Production in Japan began in 1970 in small volumes, and in the usual and expected pattern, exports were insignificant while domestic demand was being focused on and while cost and quality were improving to world levels. Japanese integrated-circuitry production has been growing in value terms at about 25 percent per year since volume production began, and is growing at about that rate into 1983. These value data mask a much faster increase in unit output, since prices of integrated circuits fall rapidly as yields improve with expanding volume. Japan continues to import substantial quantities of integrated circuits with import sales increasing about 16 percent per year. The result, however, is that Japanese dependence on integrated-circuit imports has been falling steadily from a 1974 high of about 30 percent of domestic demand, and continues to decline as Japanese producers increase both their share of the domestic market and increase the proportion of production being exported. Japan's trade balance for integrated circuits has swung from negative only four years ago to a strongly positive balance.

It is of special interest to note the destination of Japanese IC exports. In 1982, 38 percent of Japan's total exports of IC's went to Asian markets, while 41 percent went to the U.S. and only 16 percent to Europe. This is yet another example of the benefits to Japan's industry of the industrializing countries of Asia. As production of simpler electronic products moves offshore from Japan to developing Asian economies, those economies in turn become large markets for the more sophisticated products Japan is moving to. The cycle is beneficial to both parties.

The competition in semiconductors is

TABLE 12

TRENDS IN THE PRODUCTION COMPOSITION RATIO OF MAJOR CONSUMER ELECTRONICS EQUIPMENT



Source: Annual Data on Japan's Electronics Industry, 1983 Edition. Electronics Association of Japan, Tokyo, 1983.

Advertisement



© JAL 1983

慮

CONSIDERATION

Japan is a country where a visitor is indeed a guest. And treated as such. With special considerations you'll discover throughout your stay.

In this tradition of hospitality, Japan Air Lines extends as much consideration to business travelers preflight and postflight as inflight. JAL offers unique Executive Service that not only makes business travel more pleasant; it helps you get the most out of your stay.

BILINGUAL BUSINESS CARDS A necessity for doing business in the Orient. For a small fee, JAL can translate business cards into Japanese or Chinese and have them awaiting your arrival in Japan or Hong Kong.

HOTEL RESERVATIONS AND SPECIAL RATES Use JAL's computer system to obtain reservations, business rates and special services through JAL's Executive Hotel Service. The hotels are among the most convenient and luxurious in the Orient.

LANGUAGE CASSETTES, TRAVEL INFORMATION Enjoyable, informative books on everything from business protocol in the Orient to nightlife. Plus the JAL Jet Age Language Courses, with Chinese and Japanese cassette tapes to help overcome the language barrier.

AND MORE There are other ways to fly to the Orient. But only Japan Air Lines is prepared to offer you so much consideration even before you depart.

TO BE PREPARED IS EVERYTHING.

JAL's unique Executive Service is available to all JAL passengers. For details, call **800-835-2246 Ext. 281.** Or send us this coupon.

JAL

JAPAN AIR LINES

Japan Air Lines, P.O. Box 10618, Long Island City, New York 11101

Dear JAL: Please send me the JAL Executive Service Folder.

Name_____

Address_____

City_____ State_____ Zip_____

JAPANESE TECHNOLOGY TODAY

essentially between U.S. and Japanese firms. According to a recent report in *Business Week*, 23 May 1983, of the top 10 producers of integrated circuits in the world, four are Japanese firms. Hitachi and NEC are ranked third and fourth in volumes after Texas Instruments and IBM, with Fujitsu and Toshiba also in the top 10. Only one European firm is in this group, Philips, whose position declined from fourth to sixth in production ranking from 1979 to 1983.

As earlier with consumer electronics, and no doubt in part due to having that example in mind, the U.S. industry has for several years been seeking various forms of support and protection from the U.S. government on the grounds that the Japanese industry enjoys unfair advantages from assistance provided it by the Japanese government. In some respects the U.S. industry's concerns seem unwarranted. For some time U.S. companies have held about a two-thirds share of world production, with the Japanese share steady at perhaps one-quarter. Japan now supplies about 5 percent of U.S. demand, while the U.S. producers supply some 10 percent of Japanese demand.

However, when looked at from a U.S. perspective, some of the trade trends are alarming. *Table 14* shows U.S. and Japanese trade balances in semiconductors over the past five years, and the level of exports between the two countries. Within this short period the U.S. has moved from a modest trade surplus in semiconductors to a considerable deficit in 1982. Japan has been in steadily increasing trade surplus through the period. U.S. exports to Japan of semiconductors have not quite doubled over the five-year period, while Japanese exports to the U.S. have increased by more than four times.

There is the possibility currently that the relatively stable overall shares of U.S. and Japanese producers that has prevailed for some time on a world-market basis may be deteriorating to the disadvantage of the U.S. There is first the fact that the Japanese position in semiconductors has tended to improve relative to competitors with each succeeding generation of technology. Thus Japan's position in the 64K RAM has been stronger than in the earlier generations of memory technology. The Japanese producers have been early in announcing sampling, and initial production of the 256K RAM, the follow-on generation of technology.

In addition to this technological progress, and to a considerable degree its cause, has been the high level and growth rate of Japanese investment in this industry. The situation was reviewed recently by the Japan Economic Institute:

"U.S. semiconductor makers have to shoulder a large part of the responsibility for letting their Japanese rivals get a foothold in the American MOS memory market. When semiconductor orders hit the skids during the 1975 American recession, the industry slashed capital spending, cutting it in half according to some estimates. Companies also pared their work forces . . . . Japanese manufacturers, in contrast, continued to add capacity and invest in new equipment over the course of the recession, with an eye to remaining competitive in the computer, telecommunications and other end-equipment markets they supply as vertical producers."

This earlier strategic error was perceived by U.S. producers, who determined not to repeat it in the future, regardless of profit consequences. This resolution underestimated both the depth and duration of the recent recession, and the capacity of Japanese competitors to step up investment levels. The Japan Economic Institute reviews the situation:

"With prices firming and then turning up this year, the 64K RAM finally became a consistently profitable item for suppliers. . . . MOS memory makers in the U.S. as well as Japan will need to wring all the money they can from the chip to underwrite the huge ($50 to $100 million minimum per company) R&D and equipment costs of the 256K RAM generation.

"In fact, overall semiconductor R&D and capital spending in the U.S. has been unusually strong in recent years, even during the recession. Plant and equipment expenditures, for example, rose sharply during the late 1970's and then jumped in 1980. Determined not to repeat the mistake they made during the 1975 slump, U.S. semiconductor makers maintained 1980's high level of capital investment through 1981. They had hoped to duplicate this feat again last year, but weak prices and the recession dashed these plans. Still, capital spending was off only modestly—perhaps 10 percent—from the 1980-81 peak.

TABLE 13

BALANCE OF JAPANESE TRADE IN INTEGRATED CIRCUITS
(¥ billion)

Year	Production	Imports	Exports	Reliance on Imports (percent)	Export Ratio (percent)	Trade Balance
1970	53.2	20.6	—	27.9%	0.0%	—
1972	72.3	16.4	—	18.5	0.0	—
1974	125.5	51.1	Y6.7	30.1	5.3	–44.4
1976	197.1	62.7	22.7	26.5	11.5	–40.0
1978	281.4	61.3	52.2	21.1	18.6	–9.1
1980	570.2	108.4	182.1	21.8	31.9	73.7
1982	835.2	127.4	285.1	18.8	34.1	157.7
CAGR*	25.8%	16.4%	59.8%	—	—	—

*Compounded Annual Growth Rate
Note: First five months 1983, compared with the same period 1982: production +25.5%; imports +7.8%; exports +56.3%; exports to U.S., +68.9%.

Source: Japan Long-Term Credit Bank; Electronics Industry Association of Japan.

TABLE 14

TRADE BALANCE IN SEMICONDUCTORS
($ million US)

	United States	Japan	United States Exports to Japan	Japan Exports to United States
1978	$166.9	$91.6	$105.6	$134.1
1979	191.4	187.5	172.2	242.0
1980	136.9	469.2	151.5	397.2
1981	19.5	545.5	168.9	385.4
1982	–330.3	780.8	190.4	583.6

Source: Japan Economic Institute Report, 12 August 1983, from U.S. Department of Commerce.

Advertisement

future communications

Beyond electronics lie optoelectronics

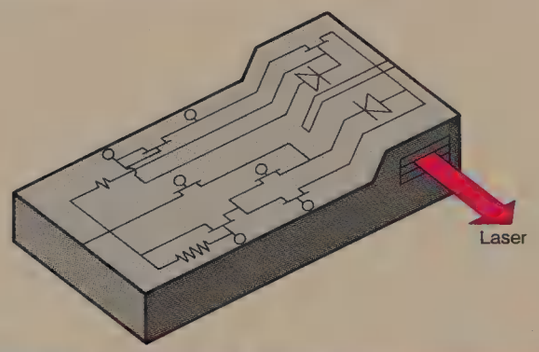
Today's digital data are usually transmitted using electronic circuits capable of frequencies of several hundred megahertz at most. But to attain the quality and reliability required of large-capacity communications in tomorrow's information society, optical transmissions in gigahertz or higher frequency bands have to be achieved. Only an optoelectronic integrated circuit (OEIC) combining a semiconductor laser or photodiode and an electronic integrated circuit (FET) would be able to provide the needed attributes. Yet, with present-day electronic circuit technology, basically a planar IC structure, it is difficult to construct both an electronic circuit and an optical device with a 3-D structure on a single substrate.

Hitachi applies micrometric skills

Using experience and know-how acquired in creating very-large-scale ICs and semiconductor lasers, Hitachi engineers have succeeded in producing a hybrid structure of optical and electronic devices on a single semi-insulating gallium/arsenic substrate. The laser diode which results emits light with an output power of several mW. And because it is completely integrated with FETs, the finished unit is compact, lightweight, and capable of high-speed gigahertz operation with improved reliability—another major breakthrough for Hitachi technology.

Limitless possibilities unveiled

This prototype optoelectronic IC is the very first important step toward realization of large-capacity communications and computers using high-speed optical processing. It can be used as a light transmitter, modulator or repeater in optical communications, or as a low-noise laser source or sensor in optoelectronic equipment. In fact, it represents an entirely new approach to handling data signals and coded information at the speed of light.



Technical excellence is embodied in all Hitachi products

The development of an optoelectronic integrated circuit is just one case demonstrating Hitachi's technological strength. You'll find other examples in all of Hitachi's products, from laser diodes and very-large-scale ICs to a whole host of high-quality electronic appliances. Our comprehensive technical expertise is your guarantee of convenience, easy operation and high reliability in every product that bears the Hitachi brand.

HITACHI

A World Leader in Technology

JAPANESE TECHNOLOGY TODAY

"Aside from worrying about financing escalating equipment and development requirements out of weak profits, U.S. semiconductor makers were looking nervously over their shoulders at what their Japanese rivals were doing. What they saw was disquieting. Like American firms, Japanese manufacturers significantly expanded capital spending during the late 1970's, and they continued to do so in both 1981 and 1982. By fiscal 1982, as a result, investment in yen terms was almost three times higher than five years before. Equally frightening to U.S. companies, Japanese semiconductor producers spent almost as much on plant and equipment in 1982 as American merchant makers did last year. Moreover, the six firms that compete in the world MOS memory market invested the equivalent of 20.3 percent of their overall sales in plant and equipment in 1982, compared with 13.7 percent in 1978, while capital spending by leading U.S. merchant companies average an estimated 15 percent of sales last year, down two points from 1981."

As the U.S. market strengthens, it is quite possible that there will be a shortage of product in 1984. This will again provide Japanese makers an opportunity to take share from capacity-constrained U.S. makers. Further, these investment rates make clear that Japan's semiconductor industry will rather soon be at least equal in size to that of the U.S. Finally these investment rates help ensure Japanese success in succeeding technologies.

The rate of technological advance in an economy is, of course, in part a function of the research capability and effort. It must be evident, however, that the application of that technology depends basically on the rate of capital investment. Invention will flow to and be applied in that economy that is fastest growing, i.e., investing most heavily in new plant and equipment. Given reasonably effective management decisions, new plant and equipment will by definition incorporate the latest and most efficient technology.

It has often been noted that in the past decades the U.K. has been an important source of significant developments in science, and has often been the initiator in the application of new technologies. However, the U.K. has for the most part during this period not been the successful site for the full application and commercialization of these technologies. No doubt the causes are many, but surely the continuing low rate of capital investment in manufacturing is significant.

The Japanese case is at the other extreme. While Japan has as yet no great distinction in discovery and invention, as the cases of consumer electronics and semiconductors suggest, Japan has been the site of successful application.

Japan's capital investment rate is exceptionally high. From 1962 to 1980, the annual increase in capital investment per employee in Japan has been about 11 percent. At the other extreme, over the same period capital investment per employee in the U.K. has been about 1 percent per employee, and it has been only slightly over 3 percent in the U.S. West Germany is intermediate, with an annual increase in capital stock per employee of about 7 percent. (See *Economic Survey of Japan 1981/1982,* p. 195.)

With the decline of the raw-material-processing industries of Japan, which are capital-intensive and had earlier drawn most of Japan's capital investment, these investment flows are increasingly into the so-called high-technology sectors, and electronics is the center of this investment flow. The capital requirements of semiconductor manufacturers are very great, as we have seen. Investment largely determines level of technology. It determines capacity levels and therefore determines the market-share winners under conditions of high demand.

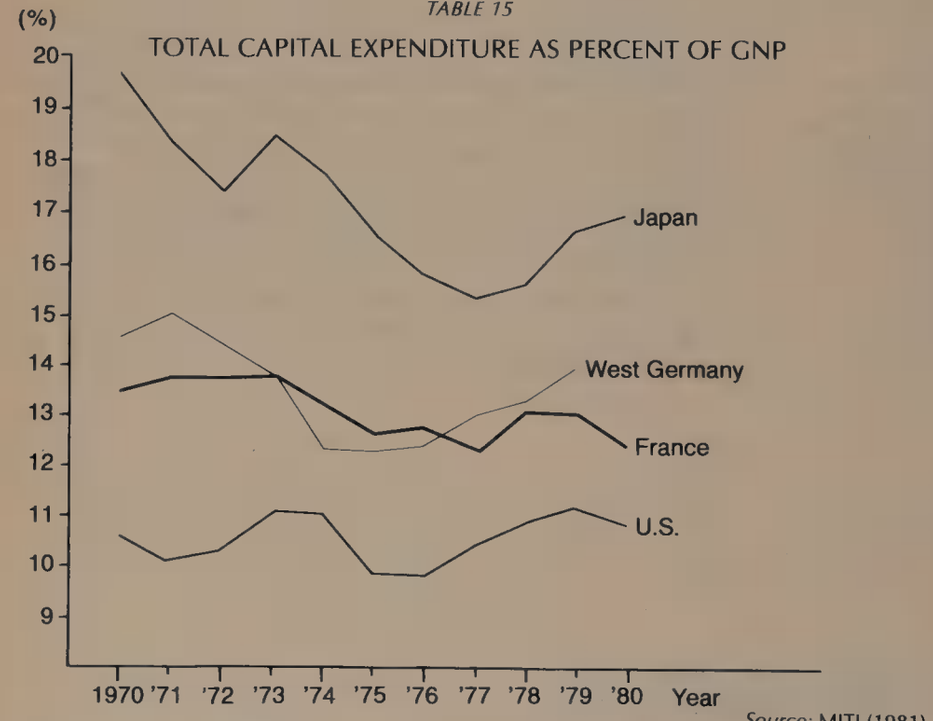
There has been a certain fashion in the U.S. to speak of "supply-side" economics, presumably a focus on savings and investment as the key to sustained economic growth in contrast to a focus on demand stimulation. It is not too much of an exaggeration to suggest that "supply-side" economics is alive, well and living in Japan. Japan's high savings rate provides the base for the kinds of investment levels shown in *Table 15*. The competitive context for Japanese business provides the stimulus to investment of these savings. The propensity for Japanese management and shareholders to take a long view of the company's future—with a focus on growth and market share—helps ensure that investment levels continue to be high, even though near-term returns would appear not to justify the investments.

From all this emerges the dilemma of the U.S. semiconductor industry when faced with Japanese competitors. The U.S. industry complains of Japanese government support and "targeting" in explaining the continuing advance of the Japanese competitors into what had been an American domain. This has given rise to much controversy about which government, the U.S. or Japanese, provides the greatest support to the industry. The argument seems largely to miss the basic issue. The Japanese government has not imposed investment targets on Japanese firms, nor has the U.S. government put constraints on the investment levels of U.S. firms. Investment is driving this process, and the differences in levels of investment between the two countries are caused by quite basic differences in how the two economies function.

This is not to say that the semiconductor race goes by default to Japanese producers. The top U.S. companies in the field are holding position, and are determined to continue to do so. It is to say however that the future of electronics is

TABLE 15

TOTAL CAPITAL EXPENDITURE AS PERCENT OF GNP



Source: MITI (1981)

Are you ready for high-performance photography?

The Minolta Program System X-700 will put you into high-performance photography!

Voted "Camera of the Year" on two continents, it can take you from full program automation to total manual control with the turn of a dial.

In the program mode, all you do is focus and shoot. Its computer balances the light for startling results. It shoots "faster" than other program cameras, so you can stop the action even in fading light.

In program Autoflash mode, you get flash performance simply unavailable in any other camera system. You literally set nothing. Minolta's advanced off-the-film-plane system accurately measures the light and controls the flash for perfect exposures.

In aperture-priority "A" mode, you select the lens opening, the computer selects the shutter speed.

And in metered-manual mode, it's total creative freedom. You tune the light to your own inner vision.

You can also expand the performance of the X-700: fire off 3.5 frames a second with the optional Minolta Motor Drive. Add the microcomputer Multi-Function Back to chronicle your life and times. And choose from more than 50 Minolta lenses and other accessories.

You get all this plus the protection of Minolta's 2-year camera/5-year lens limited U.S.A. warranty.

See the X-700 at your high-performance Minolta dealer. Take your photography into the fast lane.

U.S.A. limited warranty registration card is packaged with product. For more information on the X-700, write Minolta Corp., Dept. H, 101 Williams Drive, Ramsey, N.J. 07446. In Canada: Minolta Canada, Inc., Mississauga, Ontario, Canada L4W 1A4.

MINOLTA



X700

ONLY FROM THE MIND OF MINOLTA

Minolta X-700 shown with optional Motor Drive 1 and 50mm 1.4 Minolta MD lens.

© 1983 Minolta Corporation

TABLE 17

INDUSTRIAL ROBOT DEMAND PROJECTION
(¥ billion)

	1980	1985	1990	Annual growth rate 1985/1980	Annual growth rate 1990/1985
Domestic demand	60	200 – 250	370 – 500	30%	14%
Exports	(Negligible)	40 – 50	80 – 100	—	15%
Total	60	240 – 300	450 – 600	35%	14%
Annual production (thousand units)	19.7	43 – 55	75 – 94	—	—

Source: Japan Industrial Robot Association

no longer to be determined only in the U.S.; Japan is fully equipped to play an equal role in future developments and will do so, and efforts to obtain trade protection for the U.S. industry will not be helpful either for the U.S. or for Japan.

Semiconductors by themselves are, as steel has been, the base on which industries are to be built. It is in their applications that their value is realized. In the context of Japan today, two broad areas of technology development based on semiconductors are showing rapid development. These are the application of microelectronics to the factory and to the office, the areas of factory automation and office automation. In each of these areas technological developments in Japan are moving very rapidly indeed.

The industrial robot and its development in Japan incorporates nearly all the main themes of Japanese technological advance. Robots in the modern industrial sense were first developed in the U.S., with substantial support from U.S. government agencies interested in their applications in such areas as nuclear plants and space programs. The first robot for commercial use was shown in Japan by a U.S. firm in 1967. The initial response in Japan, as in the U.S., was limited since the machines were still crude and costly. Initial production in Japan was undertaken under license from a U.S. firm, Unimation, a pioneer in commercialization of robots.

The field developed very slowly in Japan, and it is important to note, in terms of the conventional views of Japan, that the Japanese government and its various agencies showed no interest in the product. The application of robots developed very little until the mid-1970's. It was at that time that Japanese firms came under tremendous pressure to reduce costs in every way possible. The focus of management attention in Japan shifted from dealing with fast growth to dealing with the consequences of exploding energy costs, very high inflation and a depression. It became the era of *genryo keiei*, of slimming or dieting management. The application of robots to the factories of Japan took off in that period. *(See Table 16.)*

Robots are not easy to define, in the sense that any sophisticated machine tool has some of the attributes of robots, and the definition of a robot can be quite broad or very narrow and specific. By any definition, however, Japanese industry now has in place on the factory floor a multiple of the number of robots to be found in any other economy. By one recent survey, using a very demanding definition, Japan has in operation about twice as many robots as does the U.S., or four times as many on a per capita basis.

The field of robotics has exploded in Japan in recent years. It is estimated that 190 firms are in robot production. Fundamental studies of robotics are being undertaken in 85 universities and public institutions across the country, the number of participating institutions having doubled in the last six years. The main themes of robot R&D are increasing speed, size reduction, computer control, weight reduction and modular design. The Long-Term Credit Bank concludes that "assembly robots that are capable of performing sophisticated functions will, in all likelihood, come into widespread use by 1985."

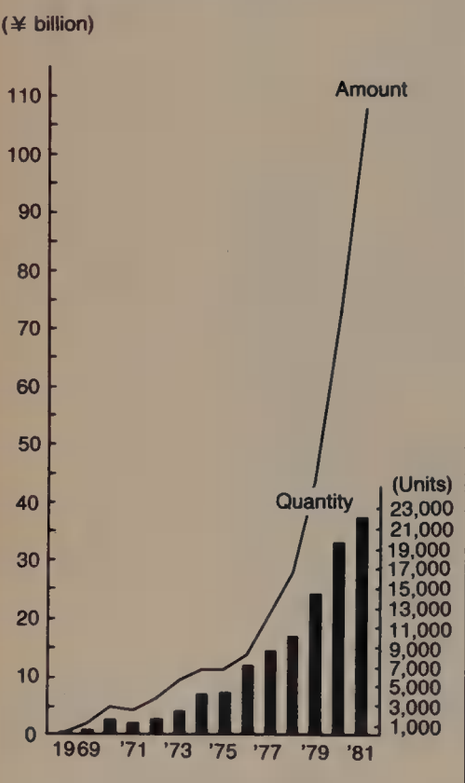
Clearly, this development is driven by the same factor that is driving Japanese progress in semiconductors, the high level of capital investment in Japan. It is not surprising to find that Japan's electrical and electronics companies are at the forefront both of development and application of robots. The high rates of capital investment in new plant and equipment provide the opportunity to set up new production facilities with the latest in all forms of automation, including robots.

Robots are expected to become a big business in Japan. *(See Table 17.)* The robot market is expected to reach about $1 billion by 1985, and to double again in the next five years. Unit output is expected to increase by some four times over the decade of the 1980's. These estimates may well be too low, as some manufacturers are talking of having tens of thousands of robots in place in a single company by the end of the decade. In any event, increased volumes will reduce costs and prices, and functions can be expected to be added in parallel. This seems likely to be a product area in which the considerable lead Japan now holds over other economies is likely to be maintained or to increase. Japanese capital investment levels alone virtually guarantee this outcome.

Just as robots are typical in being an improved-on imported technology, and in being driven in development by Japan's investment rates, so the field is representative of the responses of the Japanese government. It was only well after the initial surge of interest and growth of the robot field that the Japanese government became involved. There was no special prescience shown by the government. It is sometimes suggested that the government of Japan has shown unusual skills in "picking winners," products, companies and

TABLE 16

JAPANESE PRODUCTION OF INDUSTRIAL ROBOTS



Source: Japan Industrial Robot Association

NEC is the title sponsor of the Davis Cup and Federation Cup worldwide.

C&C
COMPUTER AND COMMUNICATIONS
TECHNOLOGY

How NEC helps curb rising costs

At NEC, we have always believed that automation effectively raises productivity and lowers business costs.

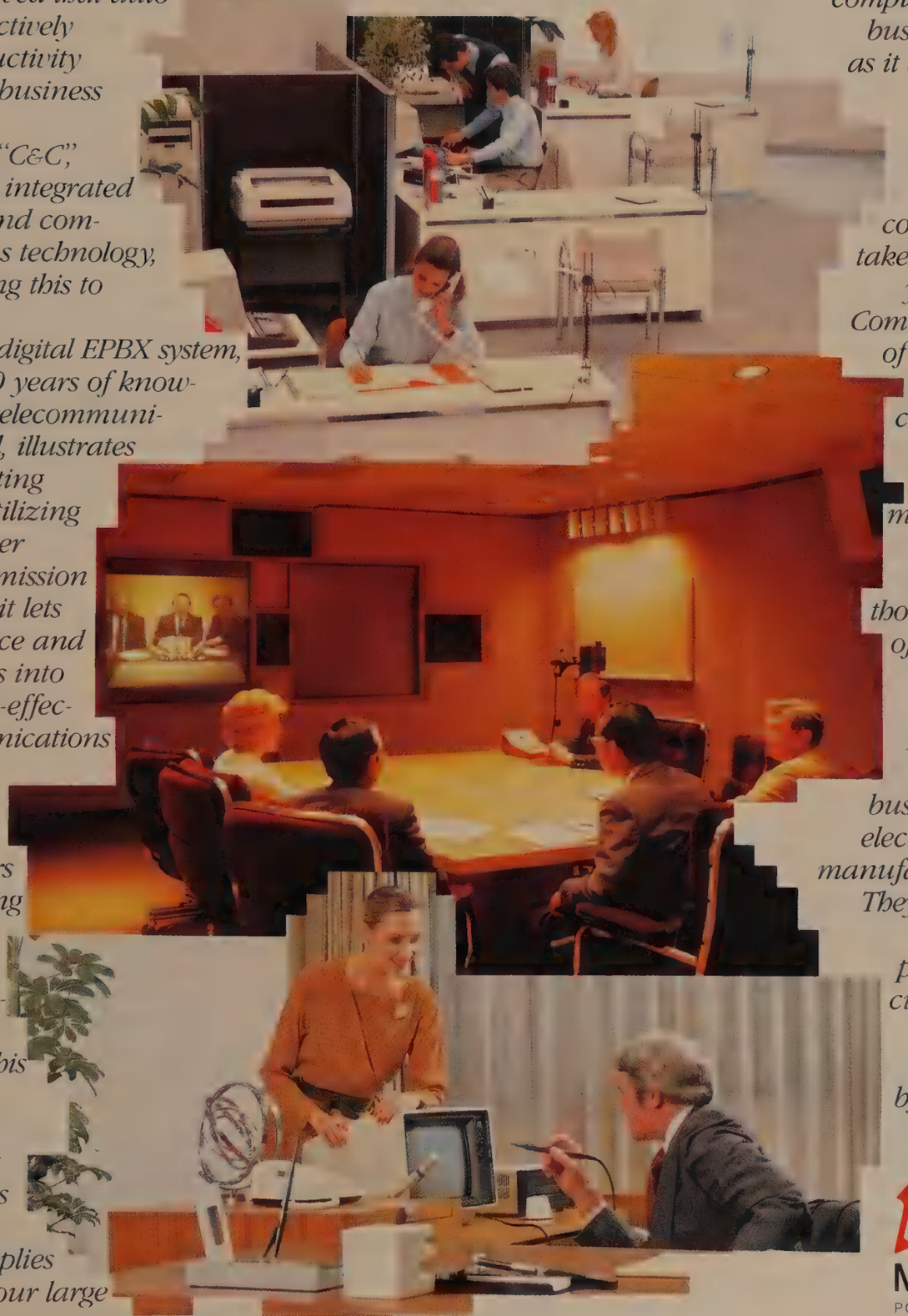
Through "C&C", our unique integrated computer and communications technology, we're proving this to be true.

The NEC digital EPBX system, based on 80 years of know-how in the telecommunications field, illustrates this cost-cutting potential. Utilizing the latest fiber optics transmission techniques, it lets you link voice and data systems into a single cost-effective communications network accessible to all.

Computers incorporating our own advanced semiconductors plug nicely into this network, trimming profit-eating expenditures all along the way. This applies as much to our large general-purpose computers and ASTRA business computers as it does to personal computers like PC6000, 8800, and 16-bit APC.

The NEC teleconference system takes our philosophy yet another step. Combining a variety of data processing systems with communications systems, it can save businesses millions of dollars annually—and business people thousands of hours of needless travel.

These are only a few of over 15,000 different industrial, business and home electronics products manufactured by NEC. They spearhead our drive to boost productivity and curb costs by way of "C&C"—a concern shared by our customers in over 140 countries.



NEC

NEC Corporation

P.O. Box 1, Takanawa, Tokyo, Japan

industries that will provide for the future of the economy. There is very little evidence of any special skill in this regard on the part of the government of Japan. That government is, however, quite capable of mobilizing effective supporting programs when it becomes apparent that a sector is of potentially high value.

In robots it was not until 1980, well after considerable private-sector activity and developments, that government programs in support of robot development took place. These have, sensibly, focused on supporting the rapid growth of the market in Japan for robots. The Long-Term Credit Bank notes four measures to promote the industrial robot industry: (1) A 13 percent write-off of purchase price as a special depreciation; (2) Smaller companies using robots in dangerous workplaces can borrow low-interest money for their purchase; (3) Smaller companies using robots to modernize production facilities can borrow from the Medium and Small Scale Business Modernization Funds; (4) The Japan Robot Leasing Company, Ltd., was set up to lease robots drawing on capital from the Japan Development Bank and other city banks.

Robots are by no means the only sector in which electronics applications to industrial production can be made. The most pervasive application is the numerical control of machine tools. Again U.S. companies were pioneers in the field, but, as *Table 18* shows quite clearly, it is in Japan that numerically controlled machine tools have found most rapid acceptance. Through the mid-1970's production in the two countries was about equal in volume. Japan now out-produces the U.S. industry by several times. Again we are no doubt seeing the impact of relatively heavier capital investment in manufacturing.

The rates of growth in turn lead to sharply lower costs and prices for this equipment. According to the Japan Machine Tool Builders Association, in 1981 a numerically controlled lathe was priced in Japan at about $55,000, compared with a price in the U.S. of $218,000, a difference of nearly four times. Thus a cycle of increased use leads to lower prices, which in turn lead to increased use. As a result robots and numerically controlled machine tools are at a price level where they can be purchased by the large numbers of small and sub-small businesses and subcontractors that play so large a role in Japan's total production capacity.

The rapid application of electronics to the factory production system has now a parallel in Japan in the automation of the office. It has long been customary to describe the Japanese office as notably inefficient, overstaffed and poorly organized. Since office productivity is generally difficult to measure, it is rather difficult to know the extent to which this conventional view has been correct. In any event, as productivity improvements through automation have been achieved in the factory, it is natural enough that the office becomes a focus of attention. This is the more true as Japanese firms have a propensity to value information highly and as the accumulation, processing and transmitting of information becomes both more important and more difficult from the sheer volumes of information available.

The progress of office automation in Japan has been much influenced by the special requirements of the Japanese language. Effective written communication in Japanese requires the use of two syllabaries of about 50 elements in each, and in addition a minimum of about 2,000 Chinese-derived ideographs. Easy written communication would require more ideographs, to a total of perhaps 3,500. It is obvious that under these conditions an effective mechanical typewriter is simply not possible. Voice communications—meetings and telephony—and handwritten communications are required.

To a Western executive struggling from under a mountain of typewritten memoranda and other communications, the Japanese system might seem to have its merits. Some part of the seeming excess of office workers in Japan can be attributed to the constraints on communication from the complexity of the written language.

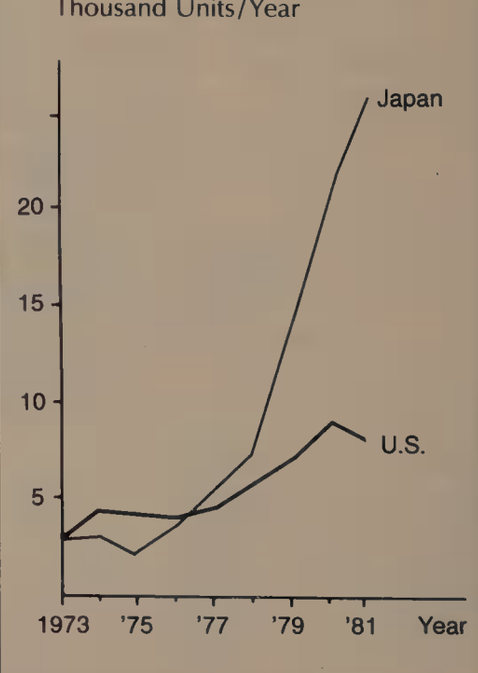
In two product areas related to office automation this disadvantage in written communications was converted to an advantage. With the problems of typing, photocopies and photo-transmittal have a high pay-off. Thus Japanese producers have become the leading suppliers of the lower-cost, lower-speed copying equipment. As Xerox aimed at the higher end of the market, holding high price levels for its products, it left an opening at the lower end of the market that Japanese producers took full advantage of. (The situation has some analogues to small cars, small farm tractors and smaller-screen television sets. As Western producers aimed their designs to attract higher prices for more complex equipment, Japanese producers focused on the lower end of the market—eventually achieving commanding positions.)

Thus Japan now produces annually some $2 billion of photocopier equipment. Moreover, the need to transmit nondigital data such as handwritten memos and the like created a very large market in Japan for facsimile equipment. The production of facsimile equipment in the world is now entirely dominated by Japanese producers. The value of production is now about $500 million and is expected to double in the next five years, with both production and exports up in 1982 over 1981 by 75 percent. Japanese equipment is now for quality, speed of transmission, reliability and price very nearly unique. Again, however, two elements that might make up part of the office of the future, the automated office, developed in Japan in order to meet the exigencies posed by Japan's written language. It was only with increased memory capacity and sharply lowered prices for memory capacity that it was possible to consider the next element of office automation, the word processor. The first Japanese-language word processor came on the market in 1978-1979. Demand has increased very rapidly. *(See Table 19.)*

Just as in the U.S., the office- and personal-computer market in Japan is in a period of very fast growth, and the word-processor capability is readily available through computer-software packages. The market for these several elements of the automated office is now some $4 billion in Japan and is expected to continue to grow very rapidly indeed. Japan, like the U.S., is fully caught up in the consequences of the microelectronics revolution.

TABLE 18

NUMERICALLY CONTROLLED MACHINE TOOL PRODUCTION



Sources: Nikkei; Japan Machine Tool Builders Association

WHEN PIONEER® LASERDISC™ CHANGED THE NATURE OF STEREO AND TELEVISION, IT BEGAN TO CHANGE THE WORLD.

Less than three years ago, Pioneer introduced LaserDisc™ brand videodisc players. Since then, they have dramatically changed the way people experience music and movies. And they've done even more. They have begun to change both education and business itself.

Very simply, the LaserDisc brand disc is a video record. A record that brings sight and sound together like they have never been experienced before.

A REVOLUTION IN RESOLUTION.

At a time when a multitude of videotape players are crowding the marketplace, the picture on a LaserDisc is astoundingly 40% better than any VCR. And even more remarkable is LaserSound. Simply stated, the stereo LaserDisc delivers is as good as any conventional recording you've ever heard.

And there's a lot to look at and listen to. Today, there are over 500 discs to choose from. Pioneer Artists offers music of every kind. Rock, pop, country and classical.

You can choose from ballet, Broadway theater and opera. And with constant participation from every major motion picture studio, the selection of movies grows bigger every day. For children, there are learning discs and games to play. Yet many people are playing games on LaserDisc brand players outside the home.

THE ARCADE INDUSTRY IS FINDING THE LASERDISC BRAND VIDEODISC PLAYER INCREDIBLY ENTERTAINING.

At the same time LaserDisc was making a great impact on home entertainment, it was changing all sorts of entertainment outside the home. Because of its ability to store and project images, it is being used to create the most exciting video games available today.

Dragon's Lair,* the first of such, is the only video game done in animation with movie-quality images. This, combined with the sophisticated strategy needed to win the game, is drawing huge crowds into arcades all over the country. And people are paying double the price to play.

And because of the interchangeability of the program disc, arcade owners should soon be able to change games without changing machines.

BIG BUSINESS IS FINDING LASERDISC A BIG ASSET.

Companies like American Bell, Inc., IBM, Apple Computers and Atari are finding LaserDisc brand players extraordinary teaching and selling tools.

American Bell Telephone employees report that training programs which use LaserDisc are more successful than those employing any other teaching medium (including videotape and classroom instruction). Learning centers combining LaserDisc players with computers are providing customers with a unique hands-on learning environment.

The scanning mechanism of LaserDisc allows you to access information anywhere on the disc in seconds. And because one can go in slow motion and even frame-by-frame, forwards and backwards, people can learn at their own speed. In addition, since the disc never wears out, the amount of learning and selling is limitless.

In fact, the next time you go shopping, you might see a LaserDisc unit set up in your department store. Some are being used as electronic catalogues and, when coupled with computers, actually sell products. Others simply help shoppers locate what they are shopping for.

THE FUTURE OF LASERDISC.

The capacity for LaserDisc to prove useful in today's society is endless. Not only have they already changed the way we listen to music and watch movies. But by providing a high caliber of audio and video input never before possible, they are quickly influencing the way we learn.

The Pioneer® LaserDisc™ brand videodisc player many people now have in their living rooms might one day be a required teaching tool in the classroom.

And in the tradition of all new technology, its application promises men and women more free time, higher levels of understanding and an incredibly exciting new way to look at and listen to the world.



PIONEER®

© 1983 Pioneer Video, Inc.

*Available through Cinematronics, Inc.

JAPANESE TECHNOLOGY TODAY

TABLE 19

PRODUCTION OF OFFICE AUTOMATION EQUIPMENT
(¥ billion)

	1975	1979	1981	1985 (forecast)
Office computer	42.7	146.3	306.7	750.0
Word processor	—	1.2	19.3	109.0
Facsimile	11.8	61.8	107.8	257.9
Copier	112.7	259.6	420.6	582.3
Total	167.2	468.9	854.4	1,699.2

Note: Office computers are small business computers priced at ¥30 million or less.

Source: Long-Term Credit Bank of Japan.

The real constraint on the application of the technologies of automation in Japan, the U.S. and elsewhere is clearly not a technological one. It is, as we have seen, in significant part an investment constraint. If savings and investment rates are high, the speed of application will increase. There is a further financial constraint, however, especially appropriate at the present time, which is that of the cost of capital. This is a subset of the savings and investment issue but is worth noting separately. To the degree that costs of capital are higher in one location than another, then investment in capital equipment will be lower. The high interest rates in the U.S. relative to rates in Japan, means that while this situation prevails, the introduction of automation will be more rapid in Japan than in the U.S.

The other major constraint on introduction of automation is that of labor relations. In the West the introduction of automation is very much slowed by the resistance of the labor force, both in the factory and office, to "labor-saving" methods. The resistance is natural enough, since introduction of automation threatens pay levels and jobs. There has been little opposition in Japan to automation. This results both from the greater job security offered employees by the Japanese company, and from the fact that compensation in Japan, through seniority pay and bonus systems, is closely tied to the success of the company. To the extent that new methods and equipment help secure the future prosperity of the company, to that extent they improve job security and income levels of the work force.

A survey by the Federation of Metal Industry Trade Unions is of interest in this connection. *(See Table 20.)* Trade-union members are reported to be positive toward or indifferent to the introduction of automation, with only a scatter of negative or critical attitudes. This is not to say that Japanese workers are not quite able and willing to be negative or critical if circumstances warrant. However, these introductions are accomplished without layoffs or firings, through retraining and transfer. The technology is not seen as a threat.

Japanese industry is no doubt much aided too in the introduction of new technologies into the workplace by the relatively high level of education of the labor force. There is no immigrant labor in the Japanese factory or office, with the problems that poses for literacy and work discipline. There is no discussion in Japan of what has come to be called functional illiteracy in the U.S.—individuals who have completed a reasonable level of formal education but who lack the ability to do simple sums or to read and write at a functional level. There are complaints about Japanese education, but these are of two kinds for the most part—first, that children are required to study too hard, not a usual complaint in other countries, and second, that Japanese universities are not sufficiently demanding, a problem indeed but not at the factory or office-work level.

Given that Japanese technology is moving forward in both office and factory automation at least as rapidly as is that technology in other countries, and given the high levels of capital investment in Japan, the equipment is becoming available rapidly for what the Japanese like to call the "new industrial revolution." This revolution is unlikely to be impeded by reactionary units of the labor force or from lack of necessary skills in the labor force. The conclusion is inescapable that automation will progress more rapidly in the Japanese economy than in other economies, and that the process will greatly increase the productivity and extend the life cycle of Japan's manufacturing sector, while moving Japan rapidly into the information-related businesses of the future.

The steel industry is both symbol and core of the first industrial revolution. As technology has changed, and as Japan and other competitors have gained cost and technology advantages over Western steel producers, the industry in Europe and the U.S. has come to be a symbol in turn of obsolete facilities, redundant personnel, protectionism and of failure to compete.

In its turn the Japanese steel industry is under pressure now from new competitors from South Korea, Taiwan and Brazil. With this, demand for steel is down. Some U.S. steel companies have responded to market pressures by retreating to diversification through acquisition of non-steel companies. Against this background of Western response to pressure it is of special interest to examine the response of the Japanese industry.

Mr. Hiroshi Kojima, Executive Vice President of Sumitomo Metal Industries, describes the Japanese steel situation:

"The first industrial revolution was led by the era of volume of material of which steel and coal were representative. In contrast, the core of the current revolution is

TABLE 20

REACTIONS OF UNION MEMBERS TO INTRODUCTION OF AUTOMATION
(number of firms reporting)

	Robots	Office Automation
Positive to introduction	51	51
Negative	2	4
Critical	3	0
Indifferent	45	54
Total firms reporting	106	109

Source: National Federation of Metal Industry Trade Unions.
From: Economic Survey of Japan 1981/1982.

Advertisement

function. Thus, if we speak of steel, we must add functional qualities to steel.

"It is not simply a matter of electronics developments. The oil shock had its influence as well. Formerly, it was possible to make money simply through setting up giant facilities, but that era is ended. Still, steel remains the basis of industry. It is inexpensive relative to its strength compared with titanium or aluminum. It remains the basis for construction. However, we cannot expect demand to grow in the future, and we are being pressed by the developing countries. Therefore, for growth, we must aim at upgrading steel. Specifically, because adding to the strength of steel is difficult, we need to make it lighter, to add materials that will strengthen it, and provide surfaces that will not rust.

"We need to develop and are now developing composite material of resins and steel. We purchase the resins, mainly from Sumitomo Chemical, but Sumitomo Metals itself is now hiring chemists, aiming at developing functional materials.

"Looking to the future, we are pursuing two approaches. One is the question of how to change our present manufacturing methods. A stable, high-efficiency system with high quality. We believe that the late 1980's will be a period of reflation for us. To this end we established last year a Facilities Technology Center for the study of new steelmaking methods. Also this past July we established a Design Control Center. This will work to develop sensors and related equipment for the automation of continuous operations. In addition, development work is proceeding on surface control and on composite materials. The shift is from volume to quality, and from volume to functional materials. In this process we must get our costs down.

"The second approach is diversification. One product is titanium alloy, initially for use in aircraft. Another area is that of chemicals from coal as oil-based by-products become expensive."

Clearly the strategy is not a retreat from steel but an expectation that Japanese producers must focus on developing those steel products that are beyond the technical capability of developing-country producers, while also looking at related technologies and products that will provide further growth.

The theme of a move to higher value added, and of the need to incorporate new technologies into mature products, is a broadly agreed-on concept among Japan's business leaders. The steel industry might be seen as a special case, but even in Japan's still-expanding and highly successful auto industry there is the view that new social and technical developments offer opportunities rather than signaling the end of an industry.

FROM VOLUME TO QUALITY . . . A MACHINE FOR SOCIAL LIVING

Masatoshi Morita, Executive Vice President of Toyota, sees the change as one from volume to quality, from mass production to more specialized, higher value added output. He states:

"In my view the 'new' of the 'new industrial revolution' is a function of the electronics boom. In addition, there has been added the new materials boom, the bio-industries boom, etc. Even with the electronics phenomenon, the real basis is from the properties of materials. It came from that kind of research. From special properties of materials came the 'new-materials' industry, and on the other hand bio-technology rose from the life sciences. This has been a process taking place in the bewilderingly short period of only six years. Thus 'technology industries' arose, seen from one point of view. Again, still differently, looking at it from the auto point of view, there has been a change in the environment for use of the auto, the result of a very complex change in society and in life patterns. User demands, or what might be called user needs, have become very diverse. At the same time domestic competition has intensified and international trade frictions have arisen.

"In all of this the former desire for expansion of volume is gone, and instead there is the problem of seeking higher value added. In the case of autos there is the question of whether this is by materials or by electronic or other means. This is the new industrial revolution, that is to say, technology. The automobile is caught up in all this. In Japan the licensed-driver population is 40 million. Life expectancy is now 74 years for males and 79 for females. Women users are increasing. Drivers are more often older persons. On the other hand, younger persons are seeking in the auto the functions of speed or of play. In these ways too the demands for autos are becoming diverse.

"Growth in volume is largely over. To realize higher value added, this increasing diversity of needs and demands must be met. To do so we must adopt the opportunities of this technical revolution.

"In this context the technical revolution has appeared. Thus in terms of making an auto there is the design issue, as well as production, supply and maintenance. It includes the problems after making the sale. In that sense the concern is with the auto before it is born and after it has died. The technical revolution appears in a number of aspects. The auto, which meets diverse interests and fits diverse objectives, also has an impact on the social economy, on trade, on social structure and as a machine for social living. We must give thought to what is called the new industrial revolution."

If there is general agreement on the broad directions that industry must pursue, toward value rather than volume, and toward full application of new technologies, there is on the part of Japan's top management, a more mixed view of the advantages and potential reach of automation. Nissan Motor's plants are a showplace for visitors to Japan who want to see robots in action. Nissan's original approach to the problem is therefore of special interest, and Executive Vice President Kaichi Kanao of Nissan provides a view of how the process was initiated:

"In the case of Japan our auto industry was late developing and we had somehow to overtake Europe and the U.S. There was a sense of danger that made for joint efforts in the company management, workers and union, and for step-by-step progress. For example, with robots it was not simply a matter of where the best place for them might be. The operators were keeping an eye out every day for where the most difficult work was. Also the Ministry of Labor conducted studies of the degree of work exposure—which tasks were the most exhausting, for example. In this way special work allowances were paid. That is, the work was especially difficult, and so special allowances were attached. Or again, for some jobs it was required that after a certain time at the job a rest period was necessary. It was in these jobs that robots were first introduced.

"From the union point of view this was an improvement of the workplace. Management was making an effort to make the workplace better. However, there was then the question of what happens to the workers involved. These people were trained for better jobs, not simply the heavy jobs of before. These included robot control and robot maintenance. This is a rather positive example of what happened.

"In addition there were men who because of their age could not do these jobs. Since a revolution cannot be made in a single stroke, they continued in their earlier work until retirement. There were also some younger men who could be better placed in quite different kinds of work. When changing jobs, of course, all levels of supervisors are different persons. These men changed jobs, but in the end there was no case of unfair handling. Everyone took notice of all this. These men changed

jobs, but in the end were not held back by that fact. Of course, these were not men who played around, and even though young, were good workers. We are a company that endeavors to have a work system and personnel system that does not cause harm to the employees. If you do not have the full efforts of the individuals, and of the company, and of the union—of all three—there cannot be a comfortable company, I believe."

Japan has moved rapidly into robotics, more rapidly by far than any other economy. Nonetheless, the economic constraints on the applications of automation and robots are still considerable.

ROBOTICS IN MASS PRODUCTION . . . RETURN ON INVESTMENT

Senior Managing Director Yoshinori Suzuki of Olympus Optical Co. discusses the issue of automation in his operations:

"We have introduced robotics into our mass production. For example, robotics is well suited to camera production, and there is an adequate return on investment. The problem is, of course, where product variety is great and where volumes are small. Where cameras are concerned, we can invest in robots and automation. However, in plating, for example, in inserting items in the plating tank, the variety of items is very great, almost infinite. Because of the various shapes of the items, the process is not mechanized.

"It is not that it could not be automated, but that the return would not warrant the investment. The process is done by people, inserting items for plating overnight and replacing them during the day. The plating charge and its control are done by computer, however. With volume products, all conveying is done automatically, but as I said, where the volumes are small it is done by hand. Still, as robots' functions are further developed, certainly they will take over the job. It is a matter of investment. The automation could be done now but it is a question of cost.

"An example is the shutter of the single-lens reflex camera. The shutter is flexible and a cord is attached. This cord must be wound on the axle, and the assembly is very complex. The work is extremely difficult, and was done in the past by young girls working most diligently. Now all of the assembly is automatic. I was trained in the old days, and did not believe it could be done by machine. But with sensors and computers, these kinds of operations are entirely possible."

Senior Managing Director of a company that represents both the surge in electronics developments and the most recent advances in factory automation, Mr. Masahide Nanbu of Toshiba notes that the unmanned factory is the concrete representation of the new industrial revolution. He notes, however: "While in general the trend is certainly in the direction of the unmanned factory, a truly unmanned plant is still some way in the future. We are involved both in unmanned-plant developments and in preparing products for sale that have application to unmanned plants. Our Nagoya plant is not unmanned but has gone far in that direction. At one of our television plants, printed-circuit-board production is unmanned.

"We have had no issue of redundancies. Staff has been transferred to other assignments. It has not been a problem. Work shifts from direct to indirect. Even in the era of automation, the need for basic technical skills does not go to zero. The staff who have developed the bases for automation have a function subsequently in building the software aspects. There is need to consider more than work methods. Worry about employment is not necessary even in an era of automation.

"Abroad, concern is voiced about robotics, but it is not warranted. It is people who make the machines work. Their intelligence is needed."

It is not surprising perhaps to find a special preoccupation with very advanced automation on the part of Fujitsu, one of Japan's leading electronics companies. Its executives propound the most elaborate conceptual approach to integrated company-wide computer-controlled operations. Executive Director Shoichi Ninomiya of Fujitsu discusses his company's current activities and plans in automation:

"What is called factory automation in our company is the integrating, in a single divisional location, of production planning, engineering, production engineering and production. Our plan is as follows. Based on office automation, we will bring together design information, production control, laboratory automation and computerized engineering, for total automation of IC design, printed boards and semiconductors. We are thinking of moving to a higher level of Flexible Manufacturing Systems. Further, this will not stop at the divisional location level. We have already completed tying sales information, production control and computers of each factory into a complete circuit. We are aiming at a system that will integrate design information.

"We have introduced robots here and there in the factories. There are two ways of thinking about robots. One thinks of the stand-alone robot, which has caused something of a fever in the world, fanned by robot specialists. The other is robots as equipment in a future automated line, integrating the robot functions. The second is the direction we are taking.

"As products become more diverse, the life cycle shortens. There would be the problem of increasing scrapping of automatic lines, and therefore a basic condition is to make them flexible and capable of producing new products.

MORE WHITE-COLLAR THAN BLUE-COLLAR WORKERS

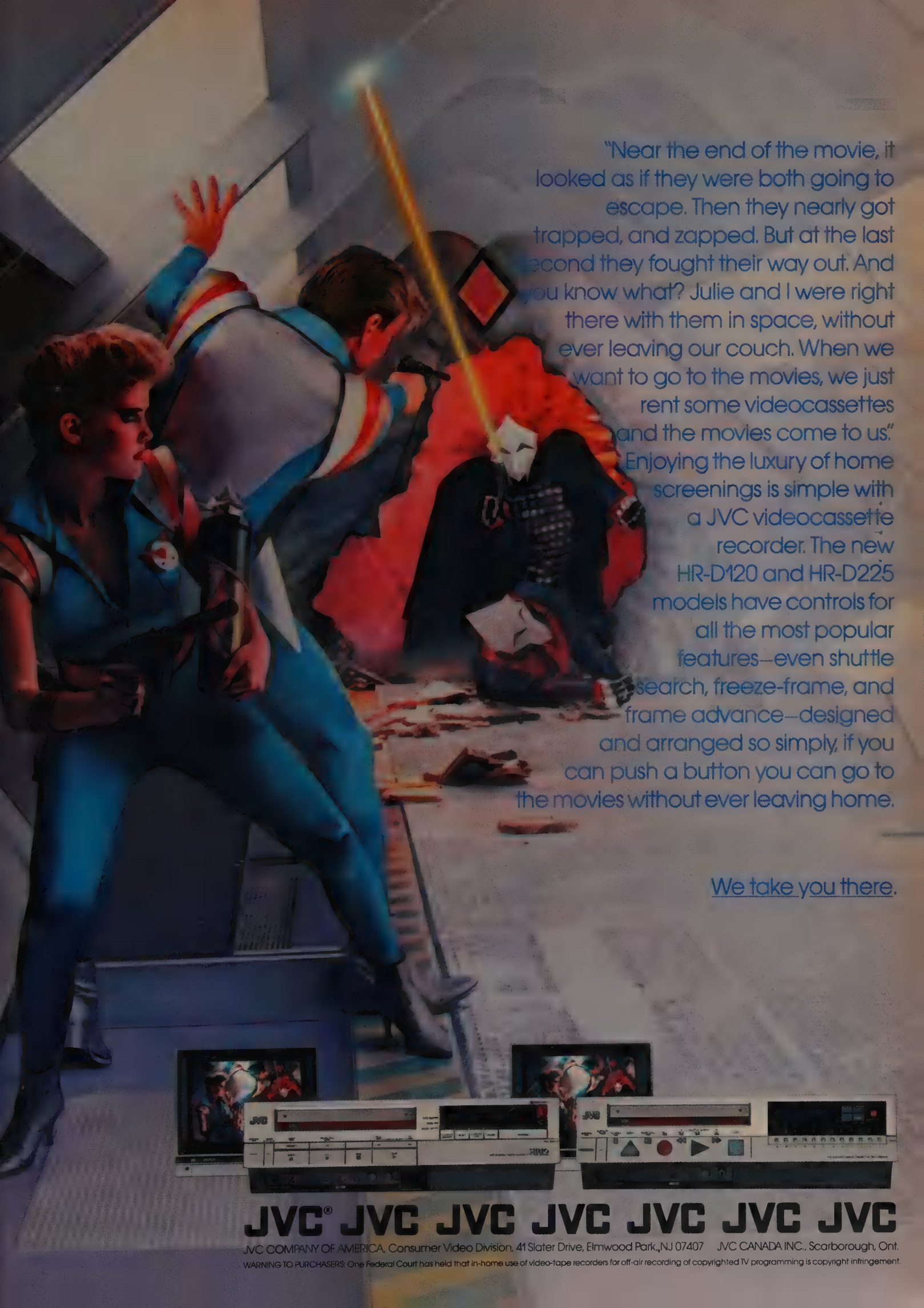
"In terms of office automation, there is office automation both in the office activities and in engineering. There is a good deal of such equipment in the company. We have over 600 personal computers in use and more than 300 word processors. The numbers are increasing annually. Currently all written material is by word processor.

"Our major planning for office automation is in design. The field might better be called automation of computer aided design. The Kawasaki plant is the main focus. Because engineers are located in the plant itself there are 180 computers there. There is perhaps one terminal for each five persons. These 180 computers at the Kawasaki plant are for hardware design only. Gradually pencils will vanish and only the word processor will be used.

"In our company there are more white-collar than blue-collar workers, so we must plan increased efficiency in the office. The Japanese language was a problem, but with Japanese word processors we can use human power in dealing with computers. The technique of converting *kana* to *kanji* is because we can use the computer. The use by computers of Japanese is becoming customary. The several difficulties of the Japanese language are disappearing."

Even in Japan, however, the enthusiasm for automation is not total. There are industries and executives who look on human skills as irreplaceable. In the view of President Shigetada Fukuoka of Nippon Kogaku, the range of production to which automation technology can be applied is still restricted:

"In our products, as in personal computers, the core is the unseen part, the very large integrated circuit, which carries the opticmechanics, the most important part. The manufacturing technology for the ULSI is not especially difficult. For example, if we produce 10,000 cameras monthly, and the factory hours are 10,000 minutes in a month, a camera is produced each minute. Thus the worker quickly becomes skillful through 10,000 repetitions in a month. However, if we are building



"Near the end of the movie, it looked as if they were both going to escape. Then they nearly got trapped, and zapped. But at the last second they fought their way out. And you know what? Julie and I were right there with them in space, without ever leaving our couch. When we want to go to the movies, we just rent some videocassettes and the movies come to us."
Enjoying the luxury of home screenings is simple with a JVC videocassette recorder. The new HR-D120 and HR-D225 models have controls for all the most popular features—even shuttle search, freeze-frame, and frame advance—designed and arranged so simply, if you can push a button you can go to the movies without ever leaving home.

We take you there.

JVC® JVC JVC JVC JVC JVC JVC

JVC COMPANY OF AMERICA, Consumer Video Division, 41 Slater Drive, Elmwood Park, NJ 07407 JVC CANADA INC., Scarborough, Ont.

WARNING TO PURCHASERS: One Federal Court has held that in-home use of video-tape recorders for off-air recording of copyrighted TV programming is copyright infringement.

equipment for making a camera, the work may take three months, and various technologies must be remembered. However clever the man is who is called for the work, it is necessary to train him in the technology. Of course, if production is increased, and the work becomes a flow, the scope for activity narrows and the amount of technology he needs to remember diminishes.

"This is to say that after the production equipment for the ULSI is produced, the ULSI production is steadily improved, but in building the ULSI machine, automation is difficult. There is much talk around of factory automation, and it would not do to miss the boat. Still, if it is something like steel, or a machine that can be slapped together, perhaps it is all right, but it will not do for precision instruments. Therefore things like robots still cannot be used. There is still the necessity for the skill of the human fingertip."

10 COPYING MACHINES FOR EACH 100 EMPLOYEES

The executive attitude toward factory automation is positive, with some reservations, and this is clearly reflected in the explosion of the market for robotics and flexible manufacturing systems. The situation with regard to office automation is rather different, reflecting in part no doubt the fact that it is only very recently that equipment appropriate to Japanese needs and to the Japanese language has become available. No doubt too many of the reservations about office automation—as well as the strong expression of need for it—would be reflected in conversations of Western executives who similarly feel a need for reduction of office overheads and who similarly do not yet see systems for office automation developed to a satisfactory level.

President Yoshihiro Mita of Mita Industrial Co. directs a leading firm in the copier field and is himself a director of efforts in office automation. Yet he too expresses some frustration over developments:

"Japan is a developed country for copiers. In Japan, for each 100 employees in a company there will be 10 machines in use, a ratio of 10 to one. In Europe and America there might be one machine for 100 employees, and they are not scattered around the offices. In Japan they are dense. This is to say that there is a considerable potential demand in the West. Canon recently has undertaken an ambitious test to develop this base. They are calling it a 'personal copier,' although whether it will be for personal use is another matter. They are already pushing it in Japan for small-scale users. If it succeeds, it will open a big market.

"The entire copy-machine industry is very active. Until recently there was little by way of smaller-size, increased functions. Now there is a sense that if the copy machine does not have these, it is not a copy machine. The market is moving to shortening the life of a machine through replacement of machines that have not been miniaturized or do not have additional functions. One could go so far as to say that a copy machine can be used for 10 or even 15 years. If the attitude can be created that one should be embarrassed to use a two- or three-year-old machine, it could increase the replacement market by several times.

"I do not really understand what is meant by office automation. The use of a copier rather than writing, of a calculator rather than doing one's own calculations, of a word processor rather than a Japanese typewriter—these working tools continue to be upgraded without limit. As a result office help can be reduced, or using the same staffing one can add to the volume of information managed.

"Information maintains the company's competitiveness. If company A manages a great deal of information, then company B does too, and their relationship doesn't change. Because the volume of information increases, people are not freed from managing data, and this causes the introduction of various machines into the office. Demand and production are born. In order to live human beings need food, clothing and shelter of some sort. More than these are non-essentials. How do we produce these non-essentials?

"Because this is the only security for a capitalist society, it is important that we think of these things. Still, in my view, office automation is proceeding extremely slowly. We still have too many people in our company."

The contrast between developments in factory automation and office automation is brought out sharply by Dr. Masaru Yamano, Executive Managing Director of Sanyo Electric:

"We have made considerable changes in our factory and office systems. Because we are a manufacturer, automation in order to get our costs down is our biggest challenge. The robot is one type of automation, as I see it, We have certainly been introducing various types of equipment. Looking at it in terms of types of products, such things as mixers, juicers, toasters and the like are older products with fixed demand levels. These are more or less necessities. Thus with these kinds of products, we can get the advantage of improved quality and lower prices through automation.

"In discussions in our industry the automation of production of goods is considered to be quite advanced. However, concerning the point of numbers of office workers in the cost structure—that is to say, overhead costs—there is a good deal of attention to the issue currently. That is, to the topic of office automation.

THE VOLUME OF PAPER REMAINS LARGE

"There is no question but that the volume of paper is very large. Even with every intention of not having right at hand those papers that are not essential, one soon has a hundred volumes. Thus I believe a paper-filing system becomes important. At the point of introducing a computer for handling these files, the current personal computer and word processor are a considerable help. However, there is still paper. Thus this is a major issue, I believe. However, I also believe it is absolutely impossible to do away with office paper. Since the era of papyrus, all management has been done through the means of paper. In order to manage volumes of important information, it is necessary to go to computer systems. Because we are in an information society, huge amounts of information have to be handled, and I believe this means that we must be able to store, retrieve and edit under any foreseeable circumstances. Still, the notion of a paperless office is doomed to failure."

President Hideo Tashima of Minolta views the office-equipment market, in contrast to the camera market, as an immature one, in which the opportunity for product development is considerable:

"Compared to cameras, I believe there is great potential for further development of office equipment. Office-machinery technology seems not yet matured. Even though various kinds of machines have been offered, there are serious service problems. While the products are maintained, the service conditions are a function of the design of the product. There has been a good deal of progress, but it is still the case that if an amateur handles the product carelessly, there can be a good deal of trouble.

"Certainly too it is the case that the price of many of these products is not high, and after-service provides income. From the supplier's point of view a good product is unreasonable. But that position cannot be defended from the point of view of the user.

"Gradually these products will develop to the point that they can be used immedi-

Advertisement

We've got the economics of copier technology down to a science.

There's never been a copier like the new Mita DC-312RE. A high performance machine with more advanced electronics and microcomputers than other copiers at twice the price. With Mita's full-featured copying system, you get practically every technological advancement available: two reduction modes, enlargement, self-diagnostic maintenance plus a 15 bin sorter and document feed. And, at 30 copies per minute, you'll spend less time copying, too. Nothing beats Mita for reliability and copy quality, consistently rated high by leading consumer reports. And with nearly a half century of experience plus a nationwide service network, Mita remains at the forefront of high technology innovation and cost-efficient performance. See the state-of-tomorrow copiers from Mita, today.

mita



DC-312 RE

Full Service Copiers mita

MITA COPYSTAR AMERICA, INC.
Heights Plaza, 777 Terrace Avenue
Hasbrouck Heights, NJ 07604
Tel: (201) 288-8900

ately after purchase without problems. It is in that sense that I speak of office equipment as being still immature. In addition, they provide only black and white.

"Providing color is very difficult. Because this equipment uses ink, it is difficult to change colors. They can be developed gradually by adding first one color and then another, but it is an interesting product area."

FROM IMPORTED TECHNOLOGY TO SELF-SUSTAINED GROWTH

As Japanese companies draw even with the West in technology, the burden of new product development falls directly onto the Japanese companies themselves. The progression in the development of Japanese industry from the exploitation of imported technology to growth self-sustained by the independent generation of new technology is exemplified by the history of NEC, Ltd. Founded at the turn of the century as the joint venture of Japanese financial interests and International Western Electric, a subsidiary of the manufacturing arm of American Telephone and Telegraph Co., NEC had its first growth as supplier of communications equipment to Nippon Denshin Denwa Kosha, the "Bell System" of Japan. The majority shares held by International Western Electric passed to an International Telephone and Telegraph subsidiary under the terms of an antitrust decree. Those shares in time passed to Japanese hands. Under the leadership of Dr. Koji Kobayashi, Chairman, NEC has now emerged as a world enterprise in semiconductors, computers, consumer electronics, robots and in pioneering of the integration of computer and telecommunications technology.

Of his company's latest investment in factory automation, Dr. Kobayashi says: "We have just opened a new plant in Chiba (near Tokyo) that is designed to provide prototype manufacturing methods. As we move to production of a new product, the manufacturing methods will be developed at this plant. It is completely computerized, with a fiber optics network to integrate all functions in the plant's operations. As the production methods are worked out in Chiba, the actual volume production will be established in one of our main manufacturing facilities."

Advances in factory automation undertaken for its own internal "market" has given NEC significant product lines for industrial markets outside. NEC is now pressing parallel advances in office automation. The Tokyo office of Mr. Toshio Ono, the company's General Manager, is directly tied to his office at the new Chiba plant. He is able to open his mail, direct it to the proper person—or throw it out, file it, or answer it—from the computer console in Tokyo tied to and displaying his desk in Chiba. At a single touch to the keyboard of his computer work station, Mr. Ono can summon any category of information a corporate officer can require. He can for example key in to a presentation of the day's news relevant to him and to the company, scan reports prepared by company staff, and make hard copy of items that he might wish to refer to later. Still something of a showplace, the equipment provides a clear indication of where NEC sees office automation and computer/telecommunications integration moving in the future.

Not surprisingly, many of Japan's top business leaders have given a good deal of thought to the question of invention and innovation. Chairman Hideo Sugiura of Honda accepts that the process of development as a follower, drawing on the developments of others, is basically different from the process of innovative development. He sees in the breadth of application of current developments a compensating balance to the fact that true innovations are less frequent today.

QUANTITATIVE CHANGE BECOMES QUALITATIVE CHANGE

"If development is late, it can be broadened very rapidly. Earlier, research work, applied research, development, application and variations could be done in one step. That is no longer possible. For example, there is an enormous quality difference between the 8K and 256K. But from 8 to 16, from 16 to 32, from 64 to 256 was only a difference in quantity. This quantitative change can bring about an enormous qualitative change. Five or six years ago it was said that the age of innovation is over. I have a negative reaction to that. Certainly it is different now from the time that Edison could make a sudden invention. But in various places, various needs and various seeds are located. At some time, on some day, when the fireworks have been placed, a chain reaction will occur and a new thing will appear. Is it a truly qualitatively new thing? Whether it is or not, the point is that from the transistor to the IC was not a qualitative difference. From the IC to the ULSI is a quantitative change that becomes a qualitative change because of the great breadth of application.

"The earlier industrial revolution was a one-way street. Now the contents of development are broad, and from the user's side the applications are broad. From both sides it has become a mutual traffic in all directions. That seems to me to be the great difference between the present and the past."

As Japanese firms gain experience in new-product development and introduction, their sensitivity to the demands of the marketplace in terms of adapting products to users' needs has become acute. Pioneer has had an enormous success especially in the field of audio equipment. It has also had some difficulties with the introduction to the market of the videodisk machine. Dr. Takeo Yamamoto, Managing Director of Pioneer, draws some conclusions from Pioneer's experience:

"There is much discussion of 'new media' as a basis for new industries, especially among engineers who are concerned with hardware. But it is clear that it will be some time before these new industries emerge. In developing new products it is necessary to establish the technology that will provide the base. But on the other hand the question must be asked: What are the real needs to be met? There must be a matching of the product technology with user needs. Again, if needs are identified, there must be software to serve as an intermediary between the hardware and the users' needs.

"However much hardware is developed, the issues are how to use it and what software can be applied to it. Taking a long view, the first requirement technologically is for hard technology. In the case of the videodisk, the first stage was to conceive of the configuration of the player. With that in hand, the need was for disk technology, that is, to go from hardware to software. To actually realize a commercial product, the next step is to pay adequate attention to the software that will actually be used, that is, to move from software to software.

THE MAN-MACHINE INTERFACE

"In this process there must be careful study of the users' needs, and an effort to gradually understand those needs. If there has only been the development of the player—the hardware—and no attention has been paid to software, the product has no meaning. One can have a product, but it is a product that meets no need. The interaction between needs and products is a seesaw game."

Despite the challenges that product innovation presents, the confidence of Japanese executives in their companies' capacity to accomplish major technological innovations is high. Dr. Fusao Mori, former Managing Director of Mitsubishi Elec-



Introducing SEIKO SPORT•TECH

Seiko welcomes you to the 21st century. With the TV Watch. The training watch for the serious runner. A talking watch that records and plays back messages. And the world's first analog quartz chronograph. See these and other Seiko wonders in the futuristic Seiko Sport • Tech Collection. You get the best of Seiko only where you see this sign.

SEIKO
AUTHORIZED DEALER

SEIKO

Setting the standard for the world, for the future.

tric, reviewing his and others' efforts in computer developments, is confident Japan will succeed in its program to develop the "fifth-generation computer."

"The biggest problem when I was in the research area concerned what is called pattern recognition, the technology of the interface with the machine. This is not the problem of output from the machine, but of input—the effort to bring together voice, letters and patterns. That is difficult. I think the technology of entering the computer from humans—the man-machine interface—became important from that time. There was also the concern of not being able to meet the competition from IBM. Our goal was to compete, because however we tried, it was not enough. The man-machine interface was the most important and difficult issue.

"The 10-year pattern recognition project was the first research effort. There is no solution with the present computers, using von Neumann methods. There has been a second- and third-stage effort, but the outlook presents many difficulties.

"We may finally be seeing a path ahead. Concerning the 'fifth-generation computer' one cannot say yet whether it will succeed or not, but I believe the value of the personnel assigned is high enough. There is a plan for a prototype in 10 years. In terms of research methods, if there is an understanding that there must be a breakthrough of the wall presented by the von Neumann methods, there should be a sufficient effort to achieve the breakthrough. There are many approaches to research, and it is important to maintain an optimistic view of what can be done. I believe that sufficient effort is being made to allow the Japanese to make the breakthrough."

As Japan has moved to product innovation, its export successes with those products have created substantial trade tensions. Japan Victor is an outstanding innovator in consumer electronics, and leader in the videotape-recorder field. The company's export success is outstanding. It is therefore of special interest to note that Mr. Toshiya Inoue, Senior Managing Director of Victor, feels that the future lies more with trade in software than with trade in hardware. He is further concerned with the present constraints in Japan on software from the still-limited development in Japan of alternative media such as cable television. He is confident this situation will soon change.

"The world is in an era of technological development, and various new products are appearing. From the time these products appear, thinking internationally, that technology will be transferred. A country will select a suitable product, build factories, hire people and make profit in the country. Whether royalties are paid does not matter. In any event we must keep this way of thinking in mind.

"In contrast, it is entirely all right to export information. That will never be a problem. That is, when a given system is in place, one country will perhaps use the software for entertainment. Again, another country will perhaps convert the information to the development of industry. Because information and software will be converted in this way, we can take the example of a computer. Its software will be appropriate to the using country. Or take the videodisk. Because it is simply a container, it will use the software that suits that country. But the hardware itself will be produced in that country.

"In the case of Japan, if the information pipeline were opened entirely and at once, there would be a big change in the country. Until now software has been tightly controlled. At about the time of the development of the VHD, there was a discussion with staff from an American broadcasting company. Broadcasts can in any event only last for 24 hours. They were concerned about videodisks and satellites. America has free competition. In a free system, good things can last, and thus there was a sense of concern and of a need to prepare for new media.

"In Japan they have been able to sit comfortably protected by controls. However, gradually CATV and the like will be coming. The broadcasting stations, the newspapers—they will face a revolution."

FROM MANPOWER, TO MACHINE POWER, TO ELECTRONICS

The advance of electronics is by no means, however, without its critics in Japan. Executive Vice President and Director Shinji Seki of Mitsubishi Motors sees problems in the continuing diffusion of electronics and automation:

"The earlier industrial revolution's origins, for example Watt's development of the steam engine, came I think with no consciousness of an industrial revolution. That is, the term industrial revolution was added by historians. The leap has been made from manpower to machine power and now to electronics. Perhaps the new industrial revolution had already begun with the invention of the diode. I feel we are already immersed in it.

"I feel there was a strong sense of thankfulness in moving from manpower to machine power, but at present? Is there a sense of thankfulness in going to the brain function of a computer in place of the human brain? I have the feeling that rather a lot of people feel frightened.

"This is a period of surplus or excess information. This is probably not a problem for people who are indifferent to information, but I think it is difficult for sensitive people. The functions of machines are surpassing that of humans. The thinking faculties of humans are, of course, clearly superior, but computers and the like are surpassing humans in reaction time and in time-sharing. This sort of thing is likely to bring about nervous problems for humans. In the first industrial revolution the machine power did not exceed the control of human power and there was an opposite effect on sensibilities.

Mr. Tozo Hikichi, General Manager, Information Systems Administration of Hitachi, Ltd., emphasizes the efforts of Hitachi to create a "computer culture." "Of perhaps 20,000 white-collar workers in our company, about half are able to write computer programs. We provide training and place importance on this not because we expect that all of these people will or should write programs in the course of their work. Rather, we think it is important that all of the staff of Hitachi have a full appreciation of what computers can do."

Japan's fast-growing, small small-computer company, Sord, and its dynamic President, Mr. Takayoshi Shiina, are often cited as evidence that entrepreneurial opportunity remains very much open in Japan today. The company got started modestly a decade ago on limited family resources; Mr. Shiina's mother remains its chief financial officer.

It has long been held that a principal obstacle to the growth of new firms in Japan is the complexity of the Japanese employment system. Employment is for the entire career, with an expectation that the employee will not leave the firm, and will not be laid off or fired during his career. In keen competition of giant firms for qualified personnel, a new small firm like Sord might well find the recruiting of such personnel virtually impossible. Asked about this barrier to the further expansion of his company, Mr. Shiina showed no concern. "I have been able to bring first-class men into Sord. You must understand that we have now some 27 subsidiaries. To attract these men, Sord must offer them the kind of opportunity that I have found. That is, they must be given a chance to run a business rather independently and given position and authority to attract them to join us."

As Japanese consider their position relative to the world there is nearly always considerable attention to those differences that may exist between Japan and the West, as sources of differences in performance. Technology is no exception, and

as Japanese executives examine relative performances they find the causes in such cultural differences as the educational system, the basic values of different religious systems and differences in the basis of the social contract.

KANJI CULTURE AND ALPHABET CULTURE

Dr. Motokazu Uchida, General Manager, Research and Development Promotion Center of Hitachi, Ltd., is of the view that differences in the education systems of the U.S. and Japan help to account for differences in software development.

"In these discussions of the relative position of Japanese technology, I believe that America is ahead in software in the sense of assembling conventional equipment and creating new functions. Also in applications software the varieties in America are much greater than in Japan. However, it is said by some that it will not be possible to catch up. That is to say, there are the differences between a *kanji* culture and an alphabet culture.

"There is the question of whether or not culture and related matters are suitable to software development. The educational system is important. There are some who argue that the Japanese educational system is still aimed at hardware. Certainly I feel it is. For example, Americans are very good at presenting and selling themselves. They have a perfect understanding of the knowhow of selling oneself. Japanese have not gotten to that point yet."

Senior Managing Director Shigeru Hayakawa of Matsushita Electric has an interesting view of the different bases for the introduction of office automation in Japan and the U.S.

"I believe that there are endemic reasons for it being difficult to introduce office automation in Japan. There also are reasons why it was absolutely necessary to introduce office automation in the U.S. The reason being that basically people are not trusted. That is to say, in a Christian country there is certainly the doctrine of original sin, of the view that human nature is evil. Thus it is a mistake to trust people. Again, it starts with the concept that people will selfishly refrain from work, that trustful relations are suspect. Thus there is the view that with office automation and factory automation, there will be fewer mistakes and selfishness will be minimized. Therefore it is easy to move to office automation.

"In contrast, in the case of Japan the ethical doctrine is that man's inborn nature is good. Thus the question in the case of Japan is: How much will office automation raise the efficiency of people? As a result, because in currently available office-automation systems the potential for improved efficiency is limited, not so many have been introduced. A large number of individual products have been brought in, but they are not yet linked together. The American system of suddenly installing a full on-line system with a whole building automated is different. We first introduce a word processor, or office computer. As usage of these increases, they gradually spread through the office. If one thinks of office automation as a total system, it seems rather easy to introduce these in America.

"In the case of Japan, rather than as total systems, office automation is entering by individual components. This seems to me a rather different way of entry than has been the case in America. For the U.S., while there is a concern with efficiency, other factors come into it. In Japan it is an issue of efficiency only. Because the use of humans in Japan is quite flexible, perhaps total systems will not spread so widely."

DIFFERENCES IN VALUES AFFECT USE OF TECHNOLOGY

Mr. Takao Nawate, Managing Director of Ricoh, believes part of Japan's less strong position in software stems from the contractual nature of Western societies and a different view in Japan of the need to compensate for services:

"I think there are differences in values between Japan and the West that affect technology advances. It concerns intangible things. In the case of Europeans and Americans, because theirs are contractual societies, naturally contracts are seen as extremely important. Even when there is no formal agreement, service is recognized as having a value, and it is well understood that something of value should not go without remuneration. However, in the case of Japan, this sort of thing is very loose. Even with computers, software may go unpaid for. It is a condition that Europeans and Americans would not be able to believe. In any event Japan is a country where being of assistance or providing service has no tradition of being tipped. I believe this difference in values is most basic."

For all of Japan's success and increasing technological independence, the spur to continued effort is still in good part the competitiveness of the Japanese company and executive, and a keen sense of the need to make further progress still. From autos to electronics the West remains the standard by which technological progress is measured. Both Mr. Kenichi Yamamoto, Senior Managing Director of Toyo Kogyo, and Mr. Susumu Aizawa, Senior Managing Director of Epson, feel their industries still learn from Western developments. Mr. Yamamoto states:

"To meet the sometimes conflicting demands for developments in autos, the basis for development will be the application of new materials and electronics to fundamental technologies. It is necessary to realize fuel savings, safety, comfort, pleasant handling and social responsibility. There are limits to the realization of all this. However, with electronic control and new materials it is possible to do the impossible. The contradictory demands must be met. The usual complaints—the ride is not comfortable, fuel consumption is high, efficiency is not good—must be done away with. This is the direction of development of the auto."

DIFFERENT APPROACH TO EACH COUNTRY

Mr. Aizawa has a world view of his company's market, yet sees the U.S. market as a testing area. He says:

"Seventy percent of our sales are overseas. We may soon encounter some troubles with the problem of exchange rates. Not all of our products can be applied to this rule, but I think it is a fundamental mistake to look on the Japanese market and overseas markets separately. In other words, what is called the domestic market is only one unit of the world market. For example, the effort suitable for the French and German markets is precisely the same as the effort required for the Japanese market. There is no reason for special handling of the Japanese market.

"Still, when one thinks of the future of world trade, with the export rate high, there are problems. In particular, when involved with the software of special-purpose computers, the approach to each country must be different. When the overseas market is in a lead position, there must be thought to how to approach the Japanese market. It is necessary to think of a U.S. approach, a British approach, a Japanese approach. There are also of course the differences between developed, newly developing and less developed markets. My point is that in the case of the developed countries, there is sometimes something of a time lag.

"For example, in the case of computers, America is four or five years ahead of Japan. This is true not just of personal computers; it is the case with all computers. The U.S. market is perhaps four times the size of the Japanese market. In these cases, there must be a different approach to dealing with the market.

"Because one can deal with the Japanese market does not mean that one can handle the U.S. market. Ultimately it is a case of learning in the advanced market. Basically it is a matter of applications software. From that consideration, the response must be different depending on the conditions of a given country."

In their overall view of the position and prospects for Japanese industry, the views of these business leaders is generally in accord with the conclusions recently presented by the Ministry of Finance's Committee for the Study of Economic Structur-

al Change and Policy. The Committee found it necessary to employ several neologisms to deal with its topic, and indeed titled its report, "A Dissertation on Softnomics," to convey its emphasis on the information industry in general and software and services in particular.

THE SOFTNOMICS OF INFORMATIONIZATION

In a partial summary of its conclusions the Committee states: "The new era is a continuation of industrialization and might be called 'the era of informationization.'

"The major direction of culture in the future, at least in the advanced economies, is not 'quantitative expansion' but is rather the 'raising of quality.' The technology that will be key is information technology. With the development of the semiconductor there has been an astonishing advance in only 30 years in the technology of information management. It has permeated all industries and even family living standards and brought about a major revolution. This trend can be called 'informationization.'

"In terms of industry this revolution has two especially striking aspects: 'the high development of basic elements' and 'systemization.'"

Whatever one may think of the aesthetics of the term "softnomics," it does convey a sense of where the Japanese see their technology and economy moving. The shift is from sheer quantity toward quality, from hardware toward software, from manufacturing toward services. It is the direction that an increasingly affluent and productive economy must move if it is to continue to progress.

With all of this affirmative view of the future, some cautionary notes may be in order. It is much in fashion currently in Japan to describe the nation's economy as being at the leading edge of technology. Yet on the whole this is not the case. In such truly frontier technologies as biogenetics, aerospace, nuclear energy and ocean-resource development the Japanese position is marginal. The strength of Japanese technology is now in the high level of development that Japanese industry has achieved in what are on the whole intermediate technologies. Japan's exceptional combination of high savings and investment, a well-educated and well-motivated labor force, an intensely competitive domestic economy and a need to seek economic security have combined to bring Japanese industry to the very front edge of conventional technologies. The issue of Japan's success at the frontiers remains to be resolved.

The publishers, authors and Ted Bates Advertising/New York acknowledge the important contributions by the members of the international business community whose messages appear on these pages:

American Honda Motor Co., Inc.
Canon USA, Inc.
Hitachi, Ltd.
Japan Air Lines Co., Ltd.
Matsushita Electric Industrial Co., Ltd.
Maxell Corporation
Mazda Motors of America, Inc.
Minolta Corporation
Mita Industrial Co., Ltd.
NEC Corporation
Nikon Inc.
Nissan Motors Corporation in U.S.A.
Okidata Corporation
Pioneer Electronic Corp.
Seiko Corporation of America
Sony Corporation of America
Star Micronics, Inc.
Sumitomo Metal Industries, Ltd.
Toshiba America, Inc.
Toyota Motor Sales U.S.A., Inc.
US JVC Corp.

In addition, we acknowledge the significant contributions to this section by the following distinguished business leaders who were interviewed:

Mr. S. Aizawa
Epson Corporation

Mr. S. Fukuoka
Nippon Kogaku

Dr. S. Hayakawa
Matsushita Electric

Mr. T. Hikichi
Hitachi

Mr. T. Inoue
Victor Co. of Japan

Mr. K. Kanao
Nissan Motor Co.

Dr. K. Kobayashi
NEC Corporation

Mr. H. Kojima
Sumitomo Metal Industries

Mr. Y. Mita
Mita Industrial Co.

Dr. F. Mori
Mitsubishi Electric

Mr. M. Morita
Toyota Motor Co.

Mr. M. Nanbu
Toshiba Corporation

Mr. T. Nawate
Ricoh Company

Mr. S. Ninomiya
Fujitsu

Mr. T. Ono
NEC Corporation

Mr. S. Seki
Mitsubishi Motors

Mr. T. Shiina
Sord

Mr. H. Sugiura
Honda Motor Co.

Mr. Y. Suzuki
Olympus Optical Co.

Dr. H. Tashima
Minolta Camera Co.

Dr. M. Uchida
Hitachi

Mr. K. Yamamoto
Toyo Kogyo Co.

Dr. T. Yamamoto
Pioneer Electronics

Dr. M. Yamano
Sanyo Electric

Special acknowledgment to the officers and staff of Japan Society for their contributions to "Japanese Technology Today." For more information about Japan Society write to 333 East 47th Street, New York, N.Y. 10017 or call (212) 832-1155.



CENSORSHIP IN A FREE SOCIETY. IT'S A BAD MATCH.

Censorship is the greatest tragedy in American literature. It constricts the mind, teaches fear and leaves only ignorance and ashes.

Today, all over the country, books are being banned, burned and censored. Teachers, students, librarians, and book and magazine publishers are being harassed. The attacks of these self-appointed censors are endorsed by our silence.

The freedom to read is one of our most precious rights. Do something to protect it.

Contact:
People For The American Way, P.O. Box 19900
Washington, D.C. 20036 or call 202/822-9450.

People For The American Way

A Molecular Basis of Cancer

Human cancers are initiated by oncogenes, altered versions of normal genes. In one case the critical alteration is a single point mutation that changes just one amino acid in the protein encoded by the gene

by Robert A. Weinberg

A tumor isolated from a patient is a large aggregation of cancer cells, all of them descended from a single "founder" cell. The single ancestor was once a normal cell, with a normal function in a particular tissue. Somehow it underwent a fundamental change. As a result of that change it began to divide and proliferate in response to some imperative of its own rather than in response to the external stimuli ordinarily required for cellular growth. Eventually this cell spawned the billions of similarly altered cells constituting the tumor mass.

The crucial event in the development of a cancer must be traced back to alterations of the founder cell. What happens to the cell to free it from the normal constraints on cellular growth? In the past few years some answers have begun to emerge. Cancer genes have been discovered in the chromosomes of tumor cells. These genes, often called oncogenes, represent the driving force behind the uncontrolled growth of many cancer cells. It is these genes that become activated in the conversion of the normal founder cell into a cancer cell. Once activated, the genes function continuously to direct cells toward the abnormal behavior that is called the cancerous state.

Pleiotropy

The abnormal behavior of cancer cells is characterized by many distinctive traits. Uncontrolled growth is the most obvious of them. Cancer cells often exhibit a shape that is very different from that of their normal counterparts. They fail to respect the territorial rules that confine normal cells to particular tissues. Many of them import sugar molecules at an unusually high rate. They rely to an unusual extent on anaerobic metabolism: energy-converting processes that do not depend on oxygen. The outer membrane of cancer cells is different from that of a normal cell and displays special tumor antigens giving the cells distinctive immunological properties.

That is only the beginning of a long list. The length of the list and the implied complexity of the cancer phenotype (the total set of structural and functional characteristics that define a cancer cell) raise important questions. Which of these traits are essential to the cancerous state and which are peripheral? If a cancer cell has 100 distinctive traits, is each of them acquired individually as the consequence of a discrete step in carcinogenesis? Does the cell pass through 100 stages in the course of its evolution from the normal state, with each stage signaling a different change? Or is a much simpler, "pleiotropic" mechanism at work? Perhaps a single, centrally acting cellular element is turned on and is then able to elicit simultaneously a large number of changes in the phenotype.

Fortunately the evidence points toward simplicity. An early indication came about 20 years ago from studies of several small DNA viruses that induce cancer in some animals. Marguerite Vogt and Renato Dulbecco, who were then working at the California Institute of Technology, were able for the first time to transform normal cells into tumor cells within the confines of a laboratory tissue-culture dish. They applied polyoma virus to hamster-embryo fibroblasts (connective-tissue cells) growing in Petri dishes and observed that the infected cultures yielded foci, or colonies, of altered cells that piled up on one another instead of forming flat layers only one cell deep as normal fibroblasts do. When the altered cells were inoculated into young rats, they proliferated to form tumors. Tumor cells, in other words, could be created by manipulating normal cells in a Petri dish; the normal cells could be induced to undergo "transformation" into tumor cells. No longer was the induction of cancer a mysterious process taking place only within an animal's inaccessible tissues.

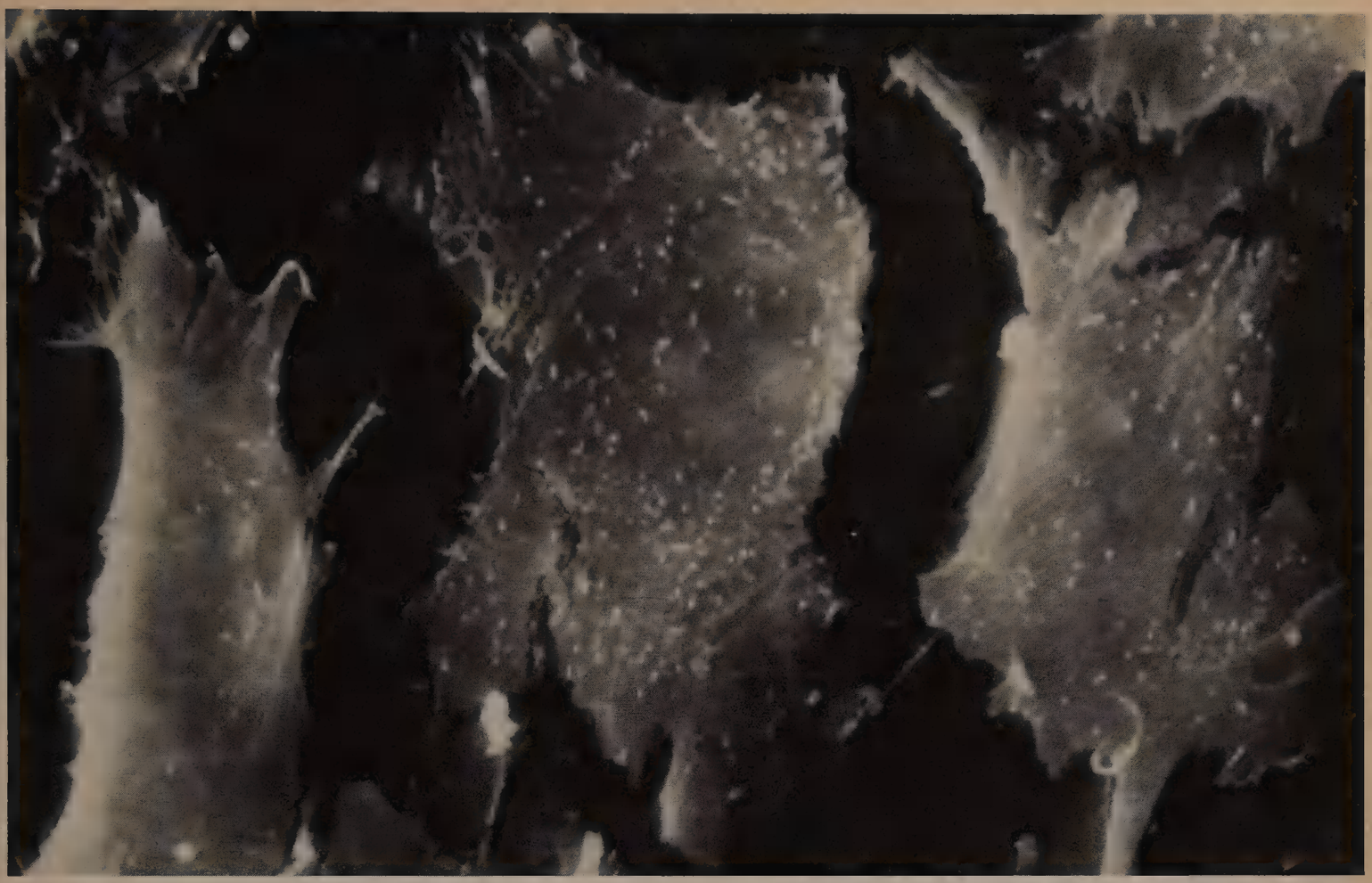
The virally transformed cells exhibited many of the altered traits associated with the cancer phenotype. Dulbecco and others showed that the transformation was somehow induced by viral genes, the segments of genetic information the virus brought into the cell in the process of infection. This information is carried by DNA molecules within the virus particles. An infecting tumor virus brings into an animal cell only a very small amount of DNA, perhaps a millionth of that which is present in the cell's own chromosomes. A very small complement of viral genes had therefore been able to elicit dozens of changes in the cells' structure and behavior. The role of pleiotropy in transformation was clearly established.

This early work led eventually to another finding bearing on the molecular basis of the transformation. If the viral genes were lost from the proliferating tumor cells or were inactivated experimentally, the cells reverted to a normal state. Not only were the viral genes required to initiate the process of transformation; their continued presence and activity were necessary to maintain the tumor phenotype.

A Genetic Basis

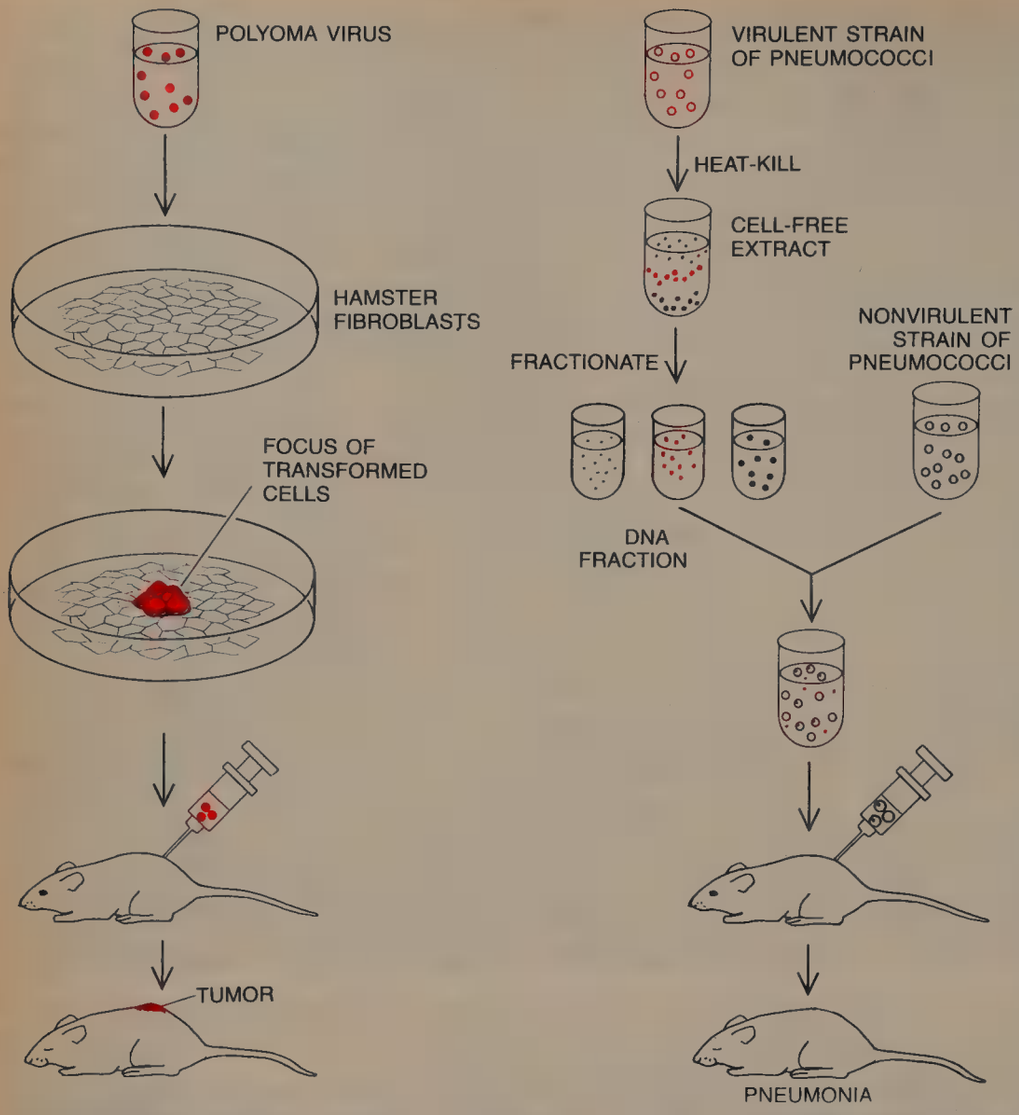
At that stage in the research one might have speculated, in a leap of logic, that analogous mechanisms operate in other cancer cells, transformed by agents other than viruses. Might not all tumor cells carry a small number of genes whose continuing activity is required to orchestrate the complex behavior of a tumor cell? Such thinking implies that the basis of cancer is genetic: that the cancerous state is maintained by genes, and not by other, "epigenetic" regulatory systems affecting the cell.

Twenty years ago this idea was far from being generally accepted, even though there had long been other indications that genes are involved in cancer. Investigators had been accumulating data on nonviral carcinogens, including various forms of radiation and a wide range of chemical agents that can induce cellular transformation. Instead of bringing novel genetic information into the cell, such carcinogens act to alter





TRANSFORMATION OF NORMAL CELLS by an oncogene is demonstrated in scanning electron micrographs made by Erika A. Hartwieg and Jonathan A. King of the Massachusetts Institute of Technology. Normal mouse fibroblasts (*top*) were treated with DNA extracted from the human tumor-cell line designated *EJ,* derived from a bladder carcinoma. The tumor DNA "transfected" mouse cells, giving rise to colonies of transformed cells (*bottom*). Whereas the normal cells lie flat and keep their distance from one another, the transformed cells round up, proliferate and climb onto and over one another. The cells were fixed with glutaraldehyde and osmium, dehydrated, "critical-point" dried and thinly coated with gold-palladium. Cells in both micrographs are enlarged about 2,000 diameters.



TRANSFORMATION OF CELLS BY VIRAL DNA was demonstrated (*left*) by Marguerite Vogt and Renato Dulbecco. They treated hamster fibroblasts with the polyoma virus. Only the small viral genome (the complete set of the organism's genes) enters an animal cell, but it was enough to give rise to foci of transformed cells with many altered traits. Injection of transformed cells into rats gave rise to tumors. A direct demonstration of the role of cellular DNA in transformation was still required. An approach to such a demonstration was suggested by the historic experiment of Oswald T. Avery and his colleagues that in 1944 defined DNA as the genetic material (*right*). A cell-free extract of virulent pneumococci was fractionated, and a fraction consisting of DNA alone was shown to render a previously nonvirulent strain virulent.

existing genetic information: the genes of the target cell. Many carcinogens act by damaging DNA and thus causing gene mutations. Once again a central role for genes in carcinogenesis was suggested.

"Suggested" is all one can say. Direct evidence was lacking. An experiment was required to demonstrate directly the presence in the cancer cell of altered genes responsible for orchestrating the cell's aberrant behavior. Such a demonstration required analysis of the cell's very complex DNA, not the relatively simple DNA of a tumor virus. One way to do this was suggested by the historic experiment that 39 years ago first demonstrated the central role of DNA as a carrier of genetic information. Oswald T. Avery, Colin M. MacLeod and Maclyn McCarty of the Rockefeller Institute for Medical Research studied two strains of pneumococcus bacteria, only one of which caused fatal pneumonia in mice. What was the molecular basis of the difference between the two strains? The investigators analyzed an extract of the virulent bacteria and isolated a purified "transforming principle" that was able to convert the nonvirulent bacteria into actively pathogenic organisms. They established that the transforming principle was DNA. The transfer of DNA molecules could transmit the genetic information for virulence from one bacterial cell to another. DNA was shown to be the carrier of genetic information specifying a discrete trait.

Gene Transfer

Five years ago Chiaho Shih undertook an analogous gene-transfer experiment in my laboratory at the Massachusetts Institute of Technology. He asked: Could the cancer trait be transmitted from one mammalian cell to another by the transfer of DNA molecules? A way to transfer genes between mammalian cells had been developed a few years before by workers in the Netherlands. Following their procedure, one traps DNA molecules in crystals of calcium phosphate. The crystals are then applied to cultured cells and act to facilitate the DNA's entry into the cells. Once inside the recipient cells, the transferred DNA becomes integrated into the chromosomal DNA.

Shih extracted DNA from tumor cells, in this case mouse fibroblasts that had been transformed into a tumor-cell line by exposure to the chemical carcinogen methylcholanthrene. The donor-cell DNA was coprecipitated with calcium phosphate and applied to cultures of untransformed mouse fibroblasts designated *NIH*3*T*3. Two weeks after the donor DNA was applied, foci of transformed cells appeared in the *NIH*3*T*3 cultures. Under the microscope the transformed colonies looked very much like those induced by tumor-virus infection. When the transformed cells were inoculated into mice, tumors grew.

The lesson was clear. The information for being a tumor cell had been transferred from one cell to another by DNA molecules. The expression of at least some part of the cancer phenotype could be traced directly to information carried in the DNA of mammalian cells that had been transformed by a chemical carcinogen. When the DNA of normal cells was similarly transferred, it had no such effect. This showed that DNA sequences in the donor tumor cells differed from analogous sequences present in the DNA of normal cells. We suspected that these differences were due to mutation of the donor tumor-cell DNA by methylcholanthrene.

The generality of those first observations could be questioned because the donor DNA had been taken from chemically transformed mouse fibroblasts and introduced into untransformed cells of the same type. It was possible that the DNA-mediated transformation we observed was characteristic only of donor and recipient cells of this special type, and not of other types of cells. That turned out not to be the case. Workers in our laboratory and in several others repeated the initial experiment with DNA's taken from a wide variety of tumor cells and found the same transforming activity. Among the human tumors that served as DNA donors were carcinomas of the bladder, colon and lung as well as fibrosarcomas, neuroblastomas and even leukemias.

It appeared, then, that some common feature must link the chemically transformed mouse fibroblasts with the cells of all these spontaneous human tumors, since the DNA of all of them has the same effect when it is transferred. Moreover, if the DNA of a human colon carcinoma could transform mouse fibroblasts, it was clear that the transforming sequences of the tumor DNA can func-

tion in cells that differ both in their species of origin and in their tissue of origin. Finally, the transforming DNA introduced into cells by gene transfer seemed to act pleiotropically, because the cells transformed by the introduced DNA now exhibited a number of traits characteristic of tumor cells. A limited amount of donor genetic information was shown to elicit several different behavioral responses in recipient cells, just as in the case of the polyoma-virus DNA.

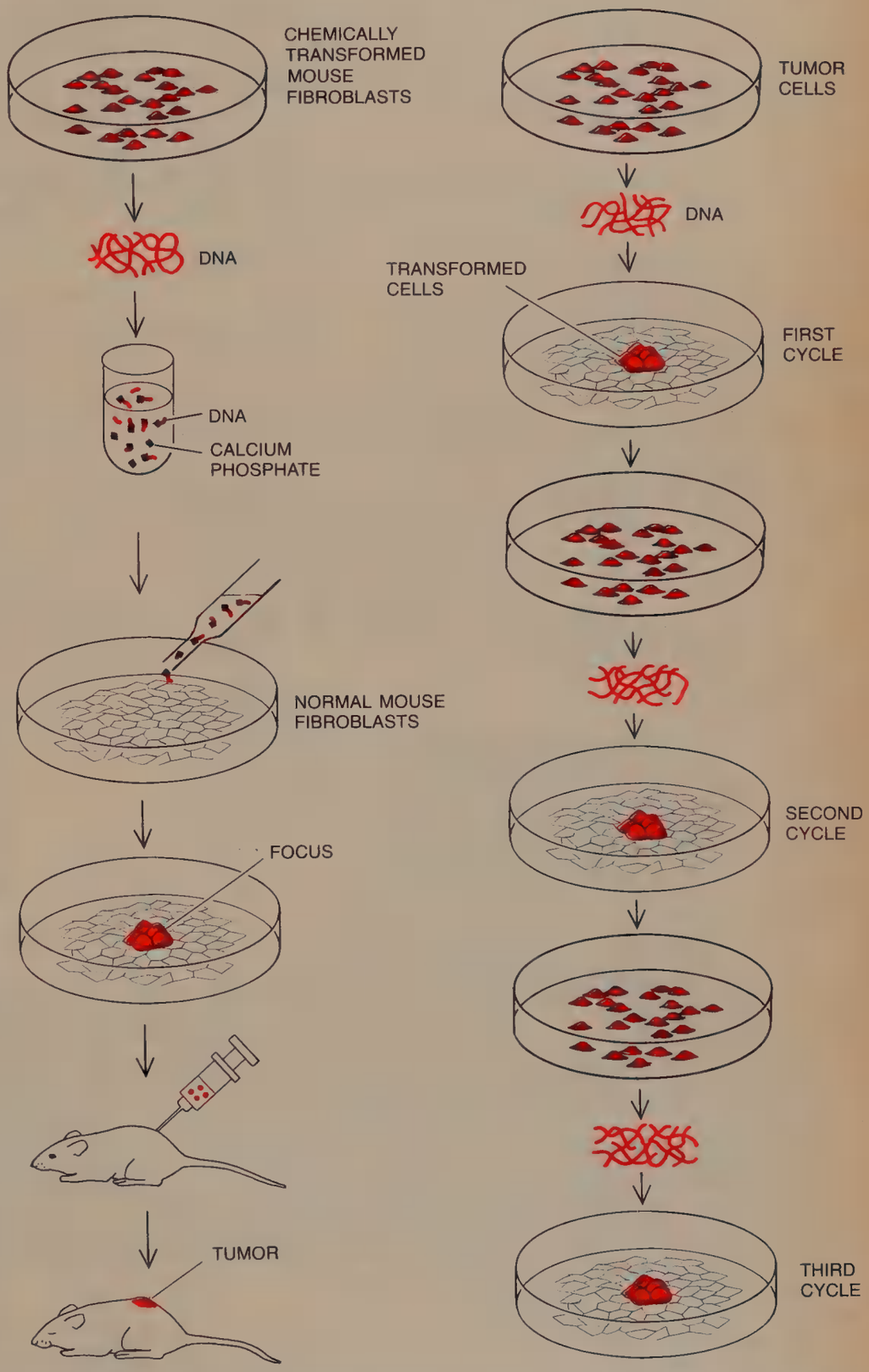
Defining the Transforming Principle

The gene-transfer experiments were informative, but further advance depended on pinning down and characterizing the specific sequence or sequences responsible for transformation. So far not even the most basic property of the tumor-DNA "transforming principle" had been established. Was it assignable to one segment of DNA or to several independent genetic elements whose cooperation was required to transform a cell? If there was indeed a single active segment, that would suggest the presence of a unique gene that functions like other discrete, definable genes in the cell's genome: its complete set of genes.

One experiment had already pointed toward localization of the transforming principle within a single DNA segment. DNA extracted from donor tumor cells and applied to mouse fibroblasts, as described above, induced several foci of transformed fibroblasts. Cells from one of the transformed foci were isolated and grown into a large culture. When DNA was extracted from this culture and applied once again to normal cells, foci of transformed cells appeared a second time. The process could be repeated through a third and even a fourth cycle of gene transfer. This suggested that the transforming principle, whatever its nature, resided in a single DNA segment whose integrity had survived the serial transfers. The successive extractions and related manipulations had the effect of breaking the long strand of cellular DNA into thousands of small segments, only a few of which could be expected to enter any single mouse cell. If the transforming activity depended on the cooperation of multiple, unlinked elements, it would surely have been lost in the course of the serial transfers since the cooperating partners would have been dissociated from one another.

The problem, then, was to find the transforming segment and study it. The design of the experiments was dictated by the details of DNA structure. The DNA molecule is a double helix composed of two strands of paired nucleotides, each characterized by one of four chemical groups called bases: adenine (*A*), guanine (*G*), thymine (*T*) and cytosine (*C*). It is the sequence of the four DNA nucleotides that carries the genetic message. The expression of this message occurs when a single DNA strand is transcribed into the nucleic acid RNA and the RNA is translated into a chain of amino acids constituting a protein. We suspected that these differences were due to mutation of the donor tumor-cell DNA by the carcinogen methylcholanthrene.

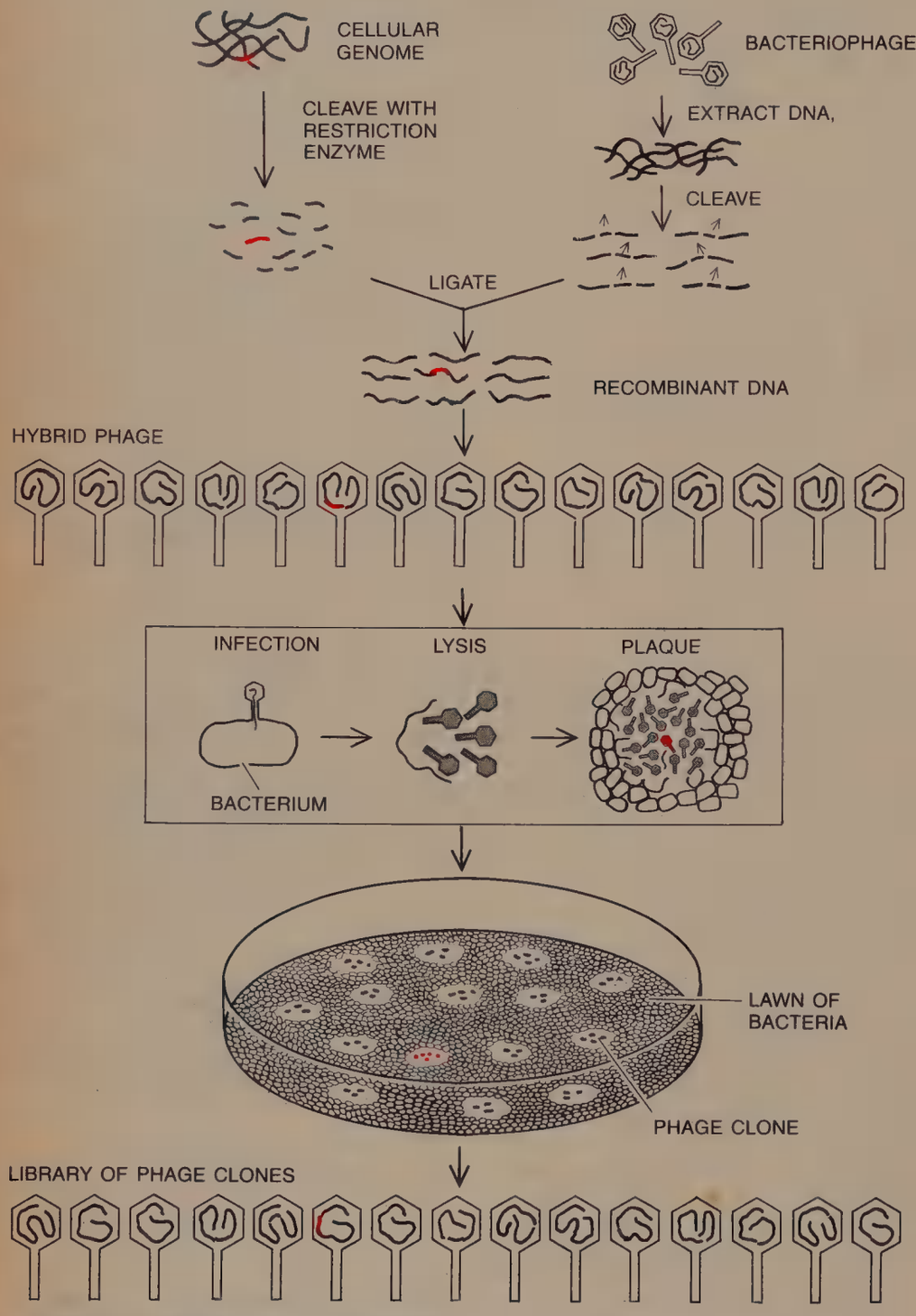
The cellular genome is composed of some six billion base pairs of DNA. An average gene is constituted from 5,000 or 10,000 base pairs of DNA. The chal-



TUMOR-CELL DNA was shown to encode cancer traits by a gene-transfer experiment (*left*). DNA was extracted from mouse cells that had been transformed by a chemical carcinogen. The DNA was coprecipitated with calcium phosphate, which facilitated its entry into normal mouse cells. Transfected normal cells gave rise to foci of transformed cells. Injection of the transformed cells into mice gave rise to a tumor. In a serial-transfer experiment (*right*) DNA extracted from human tumor cells transformed normal mouse cells. The transformed cells were grown into a large culture whose DNA was extracted in turn and again served to transform mouse cells. The process was repeated in a third cycle and even a fourth. Only short segments of DNA could survive the process of repeated extraction and coprecipitation, and so the successive transformations showed that the transforming agent must reside in a single segment.

lenge was to dissect a single genetic segment, carrying the transforming principle, away from a millionfold excess of unrelated genetic material in the cell. The newly developed techniques of gene cloning made it possible. They enable one to retrieve a single genetic segment from the cellular genome and amplify it to make many thousands of identical copies. With this material in hand one could study a pure gene, uncontaminated by the enormous jungle of complexity that is its normal genetic environment.

The various strategies of gene cloning, some involving bacterial plasmids (small circles of extrachromosomal DNA) and some involving bacteriophages (viruses that infect bacteria), have a common logic. The cellular genome is broken into several hundred thousand segments. Each of these segments is inserted into the genome of a vector, often the small DNA genome of bacteriophage lambda, which infects the bacterium *Escherichia coli*. Each resulting hybrid phage carries a single inserted DNA segment along with some of its own genes. The collection of thousands of hybrid phages is often called a genomic library because in aggregate it carries the entire store of genetic information in the cellular genome: every part of the original genome should be present in one or more of the phages constituting the library.



GENOMIC LIBRARY can be established by cloning genes in the virus known as bacteriophage lambda. The cellular and the phage genome are cleaved with a restriction enzyme. A segment of the phage DNA is discarded and the remainder is joined to a cellular-DNA fragment with the enzyme DNA ligase. The recombinant DNA is packaged in the protein coat of phage lambda. When a hybrid phage infects the bacterium *Escherichia coli*, the phage multiplies, lysing (bursting) the cell; phage progeny go out to infect and kill nearby cells in the bacterial culture, forming a plaque, or hole. When a "lawn" of bacteria in a culture dish is infected by a large number of different hybrid phages, each plaque in the lawn is inhabited by a single clone of phages descended from the original infecting phage. Each clone carries a different fragment of cellular DNA. The problem now is to identify the clone carrying a particular gene of interest (*color*). It is ordinarily done by probing the clones with DNA or RNA known to be related to the desired gene, but no such probe was available to aid in the search for transforming DNA.

Each of the phages can be amplified independently by growth on *E. coli*. When the phages are applied to a "lawn" of *E. coli* in a culture dish, each phage infects a separate bacterial cell and proceeds to make many copies of itself. The phage progeny kill the cell and go on to infect adjacent cells, yielding thousands of descendant phages within a few hours, all of them localized within a single plaque, or hole, in the lawn of bacteria. The population of phages in a plaque is clonal: it is descended from a single ancestor. From a clone of phages one can isolate thousands of copies of the inserted DNA, and in this way one obtains a clone of the DNA segment.

Scanning a Phage Library

The central problem in such a cloning experiment is to identify those few phage clones in a library that carry the inserted DNA of interest. The most convenient way to do this is with a probe, a single-strand DNA segment whose nucleotide sequence is closely related to that of the gene to be cloned. Because the two strands of a DNA molecule are complementary (*A* pairs with *T* and *G* pairs with *C*), the probe "hybridizes" specifically with single strands of the sought-after DNA. By applying a probe labeled with a radioactive isotope to the phage plaques, one can trace and precisely localize the plaque carrying the DNA of interest. (Detection is made possible by the procedure termed autoradiography.)

The transforming sequences, however, could not be identified by this standard procedure. No sequence-specific probe was available. The DNA we sought was defined and recognizable only in terms of its biological activity: the ability to induce transformation. Novel techniques were therefore required.

Three laboratories undertook the cloning task, each with a different strategy. One group, led by Geoffrey M. Cooper of the Dana-Farber Cancer Institute, associated with the Harvard Medical School, simply did without a

The average executive spends a lot of busy hours sitting behind a desk.

While the average desk sits there doing nothing.

But equip any desk with an IBM computer or terminal—like our new Personal Computer XT—and suddenly a dull piece of furniture becomes a very smart piece of business: The Smart Desk from IBM.

The Smart Desk can give you all sorts of timely information, any time you need it.

It can produce graphs and spreadsheets. Help you compile reports. Keep your calendar, even send and receive electronic mail.

And because no one likes to wait for important facts and figures, our desktop computers are designed to connect you directly with your company's central data files and office systems. Or with the Smart Desks of managers, secretaries and other professionals around the office.

IBM Smart Desk products are easy to use, so you'll start reaping the benefits almost as soon as you sit down. And when you put yourself behind The Smart Desk, we put our famous service network behind you.

To learn more about IBM desktop computing, call your IBM representative or visit your nearest IBM Product Center.

And we'll show you a smart place to spend your career.

IBM®

The Smart Desk from IBM.

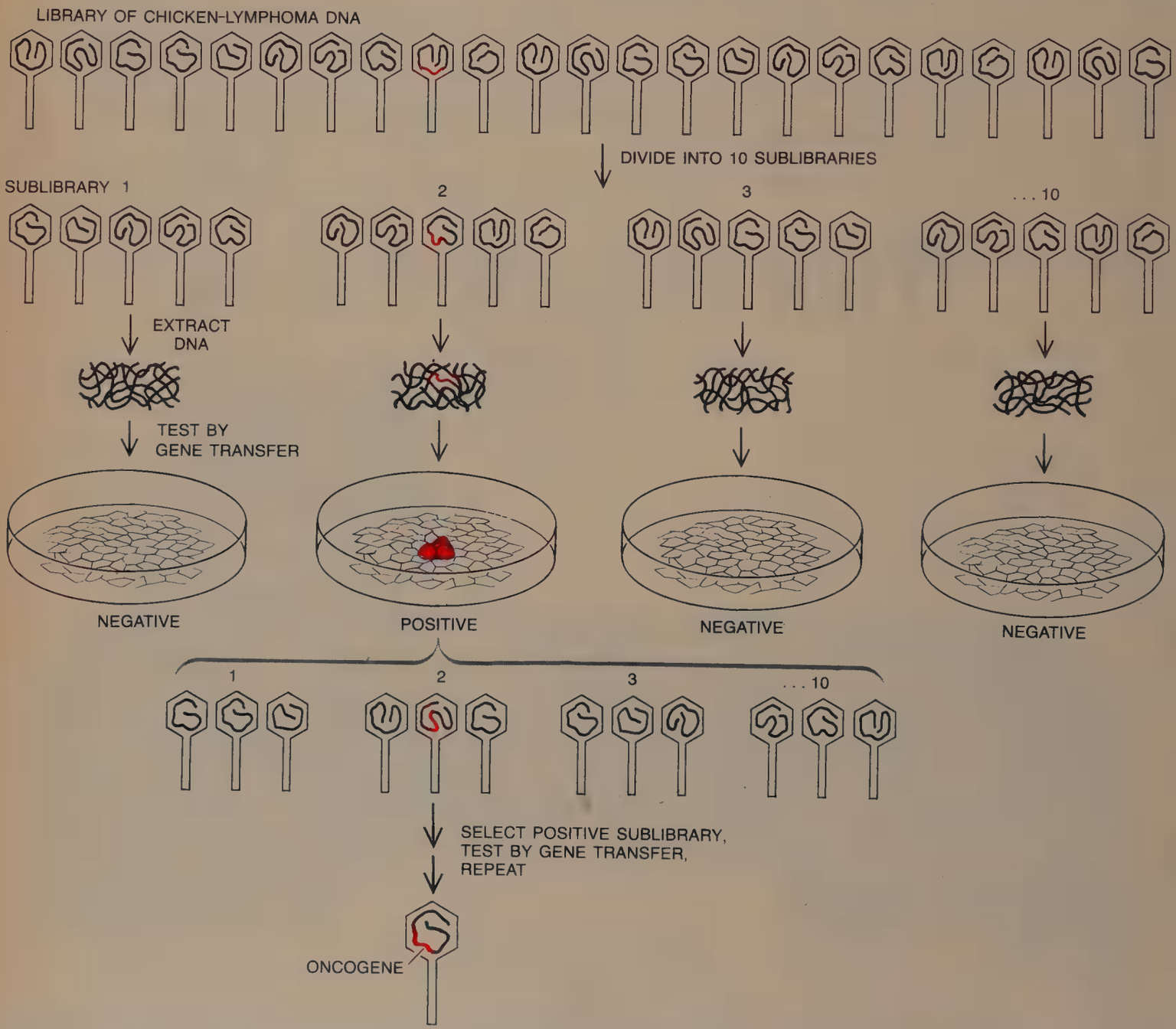


© IBM CORPORATION

probe. A genomic library was constructed from the DNA of chicken lymphoma cells, which were known from gene-transfer experiments to carry a transforming sequence. The resulting collection of several hundred thousand hybrid phages was divided into 10 sublibraries. DNA was prepared from the phages of each sublibrary and tested by gene transfer for its ability to transform mouse fibroblasts. Any positive sublibrary was further divided into 10 sub-sublibraries, each of which was screened by gene transfer. This exponentially narrowing search routine eventually led to a single phage clone carrying potent transforming activity, and thus to the transforming DNA.

An alternative approach was taken by Shih in my laboratory. He began with the DNA of a human bladder carcinoma. He transformed mouse cells with the human tumor DNA and then used the transformed mouse cells' DNA to transform a second cycle of mouse cells. The library was constructed from the DNA of these secondarily transformed cells. The DNA of the mouse cells now contained only a very small amount of human DNA: that directly associated with the transforming activity. The remaining human DNA segments had been discarded in the course of the serial gene transfer, which selected only for DNA with transforming activity. Shih pursued this small amount of DNA with a probe that would recognize only DNA of human origin: a probe for what are called *Alu* sequences. These sequences are present only in human DNA, where they are scattered randomly through the entire genome. By screening his library with the *Alu*-specific probe, Shih identified and isolated a phage clone carrying human DNA. On testing by gene transfer this human DNA was found, as it was hoped, to contain the sought-after transforming activity.

Michael Wigler's group at the Cold Spring Harbor Laboratory began by linking a bacterial marker gene to each segment of the human bladder-carcinoma DNA. By serial gene transfer they prepared a mouse DNA carrying only the human transforming DNA and a closely linked copy of the marker gene. They constructed their library by inserting the mouse DNA into special phages having defects in the sequences encoding several phage proteins. Only phages



SEARCH ROUTINE was devised by Geoffrey M. Cooper's group to isolate the transforming DNA. They constructed a library of some 200,000 phage clones from the DNA of chicken lymphoma cells. They divided the library into 10 sublibraries and tested the DNA of each sublibrary by gene transfer. If the transfer yielded transformed cells, the sublibrary involved, which necessarily included the DNA of interest, was divided in turn into 10 sub-sublibraries, each of which was tested. Positive subgroups were tested successively until a single phage clone was found that carried transforming activity. From it the chicken-lymphoma transforming gene, or oncogene, was isolated.

Human Factors Engineering. Interiors designed scientifically to minimize the possible distractions from your driving. A seat has to look great, but it has to fit great too.

It may take us two years and countless clay models to arrive at a more comfortable, durable seat for new GM cars and trucks. But we think it's worth it.

And we believe old Leonardo would have thought so, too.

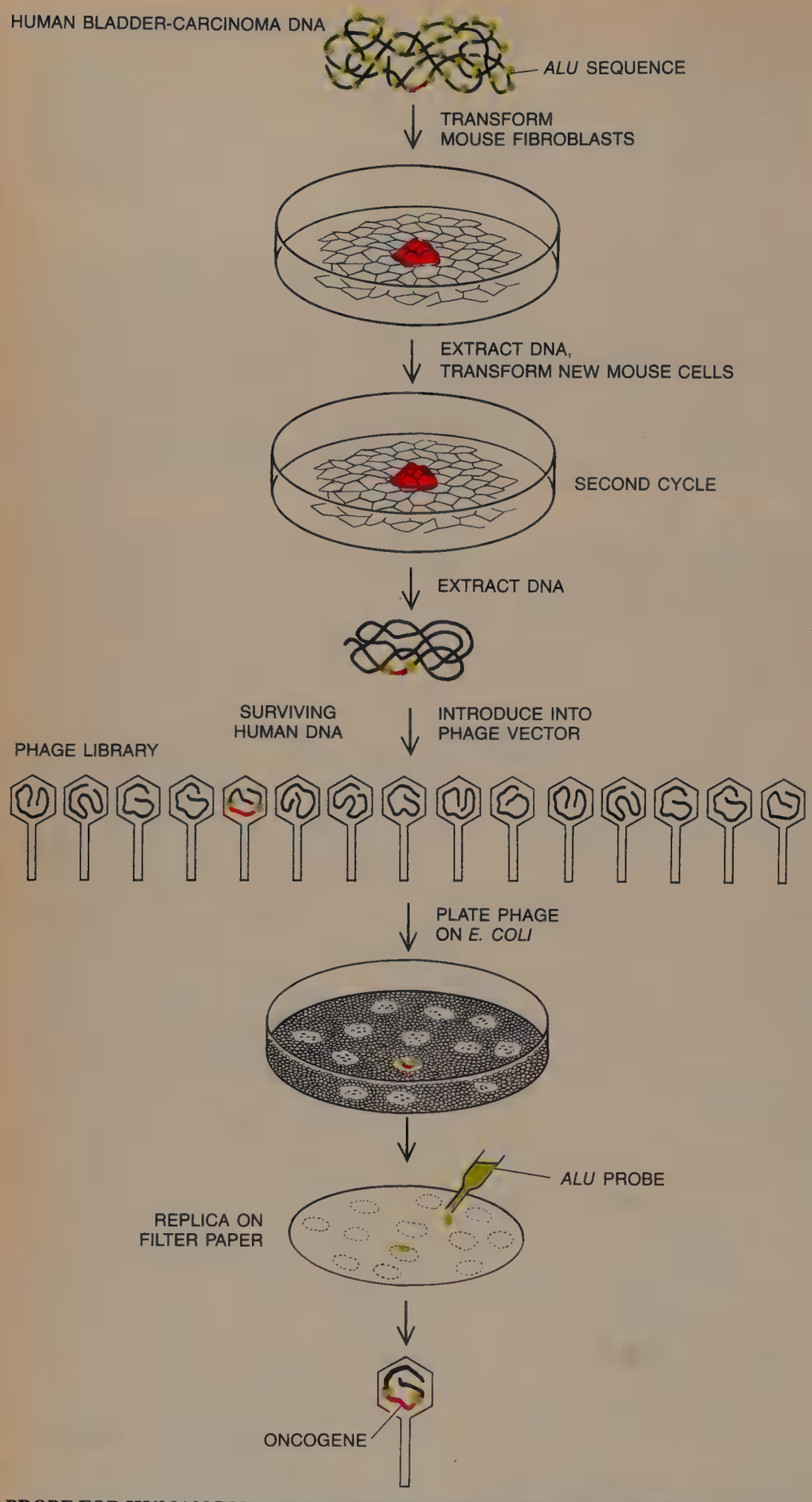
We believe in taking the extra time, giving the extra effort and paying attention to every detail. That's what it takes to provide the quality that leads more people to buy GM cars and trucks than any other kind. And why GM owners are the most loyal on the road.

That's the GM commitment to excellence.

GM



Nobody sweats the details like GM.



PROBE FOR HUMAN DNA enabled Chiaho Shih in the author's laboratory to find the transforming gene of a human bladder cancer. Distinctive short DNA sequences called *Alu* (*green*) are scattered throughout the human genome. Mouse cells were transformed with the tumor DNA; the DNA of the transformed cells was applied to transform new mouse cells in a second cycle. The DNA from the serially transformed cells could include only the small amount of human DNA that was directly associated with transformation (*red*). A phage library constructed from the DNA of the serially transformed cells was plated on *E. coli*. The DNA of the resulting plaques was transferred to a filter-paper replica, where it was probed with *Alu* sequences labeled with a radioactive isotope. Autoradiography revealed the location of the plaque carrying the DNA with which the probe hybridized. The phage clone from the plaque was tested for transforming ability and was shown to carry the bladder-cancer oncogene in its DNA.

carrying the bacterial marker gene, which counteracted the genetic defects of the crippled phage, could grow well on a bacterial culture. Among the few viable phages that grew out were several carrying the elusive transforming segment.

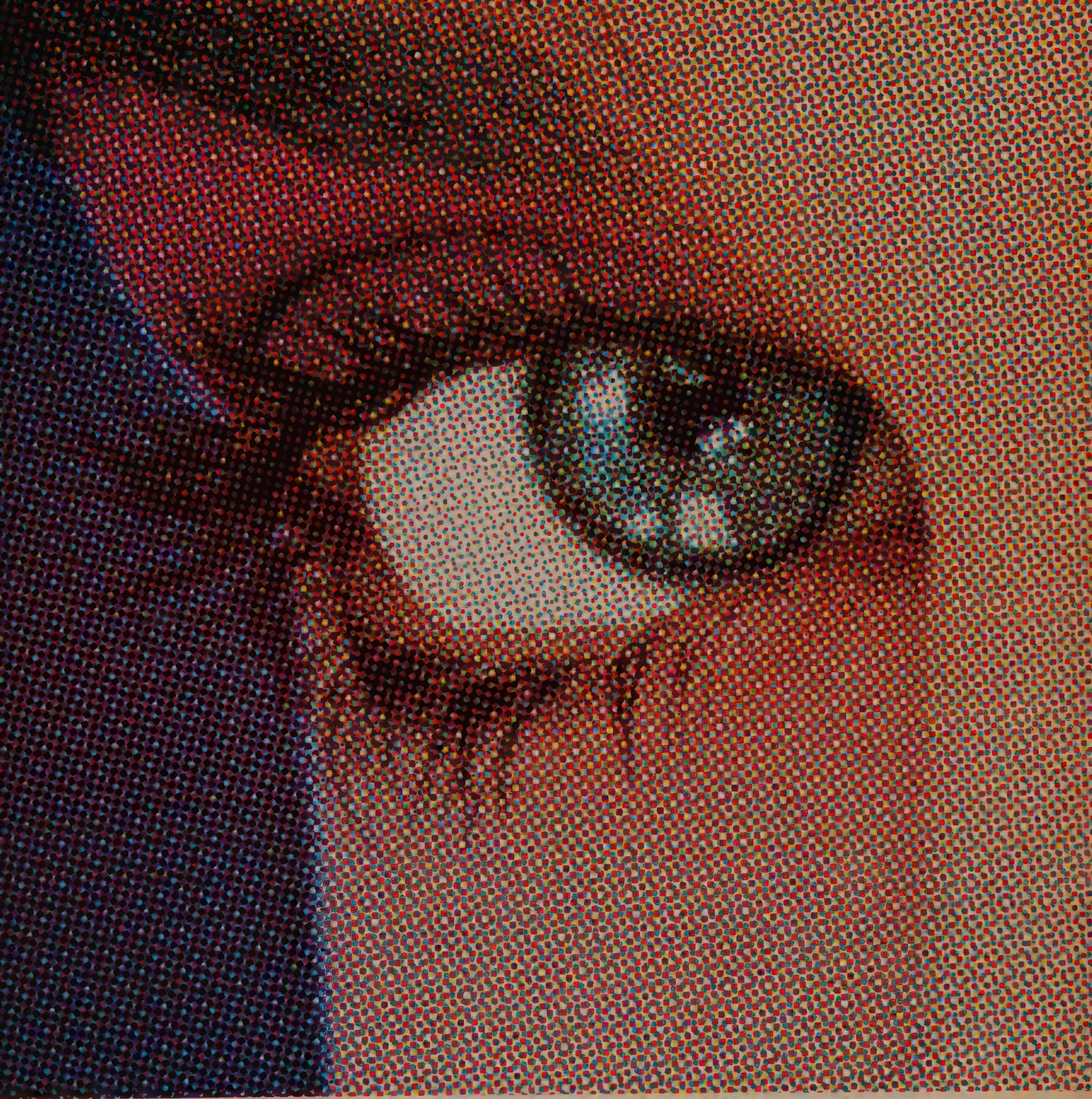
Oncogenes

The successful isolation of transforming DNA in three laboratories by three different methods directly associated transforming activity with discrete segments of DNA. No longer was it necessary to speak vaguely of "transforming principles." Each process of molecular cloning had yielded a single DNA segment carrying a single gene with a definable structure. These cloned genes had potent biological activity. Whereas two micrograms of the original bladder-carcinoma DNA had induced, on the average, one colony of transformed cells, a comparable mass of the cloned transforming gene induced as many as 50,000 foci. The transforming activity previously attributed to the tumor-cell DNA as a whole could now be assigned to a single gene. It was an oncogene: a cancer gene.

What is the origin of a human oncogene? The answer emerged, in Wigler's laboratory and my own and in Mariano Barbacid's laboratory at the National Cancer Institute, from a study of the human bladder-carcinoma oncogene that had been isolated, as described above, from a tumor-cell line variously designated *EJ* or *T*24. When the cloned oncogene was exploited as a probe for exploring the normal human genome, it hybridized readily and strongly with a related DNA sequence in the normal genome. Moreover, the oncogene and the normal gene were found to be the same size. Further characterization of the two genes showed that they were almost identical, but of course we knew they could not be absolutely identical because they functioned so differently: clones of the oncogene transformed cells, whereas clones of the related normal gene had no such effect. They could not be identical twins. Rather, it was clear that the oncogene is a slightly altered version of the normal gene, which is often called a proto-oncogene.

The precursor of the bladder-carcinoma oncogene is not the only proto-oncogene being carried around in human DNA. To date three other active oncogenes of human tumors have been traced to corresponding proto-oncogenes. Why are genes maintained in the human genome that, with slight alteration, become agents able to transform cells and generate tumors? Why should an organism carry the seeds of its own destruction?

All the answers to these important questions are not yet in, but one point is



Sometimes it's important to color the truth.

Did you ever hold a printed picture close enough to see the colors break into dots?

Most color printing is done with dots; tiny dots whose precise arrangement is extremely critical.

The 3M "Matchprint" system tells printers the truth about dot patterns with the most accurate close up available prior to printing. Corrections can be made, true colors reproduced and costly extra proof press runs reduced. Color proofing costs can be cut in half.

We answer the needs of the graphics industry with everything from photographic film to printing plates; even a cost-efficient way to buy advertising in magazines.

Listening to people's needs has helped 3M pioneer over 400 products for the communication arts field. We now make everything from lithographer's tape and overhead projectors to video recorders.

And it all began by listening.

3M hears you...

For a free brochure on all 3M products and capabilities write: Dept. 080711/3M P.O. Box 22002, Robbinsdale, MN 55422.

Name_____

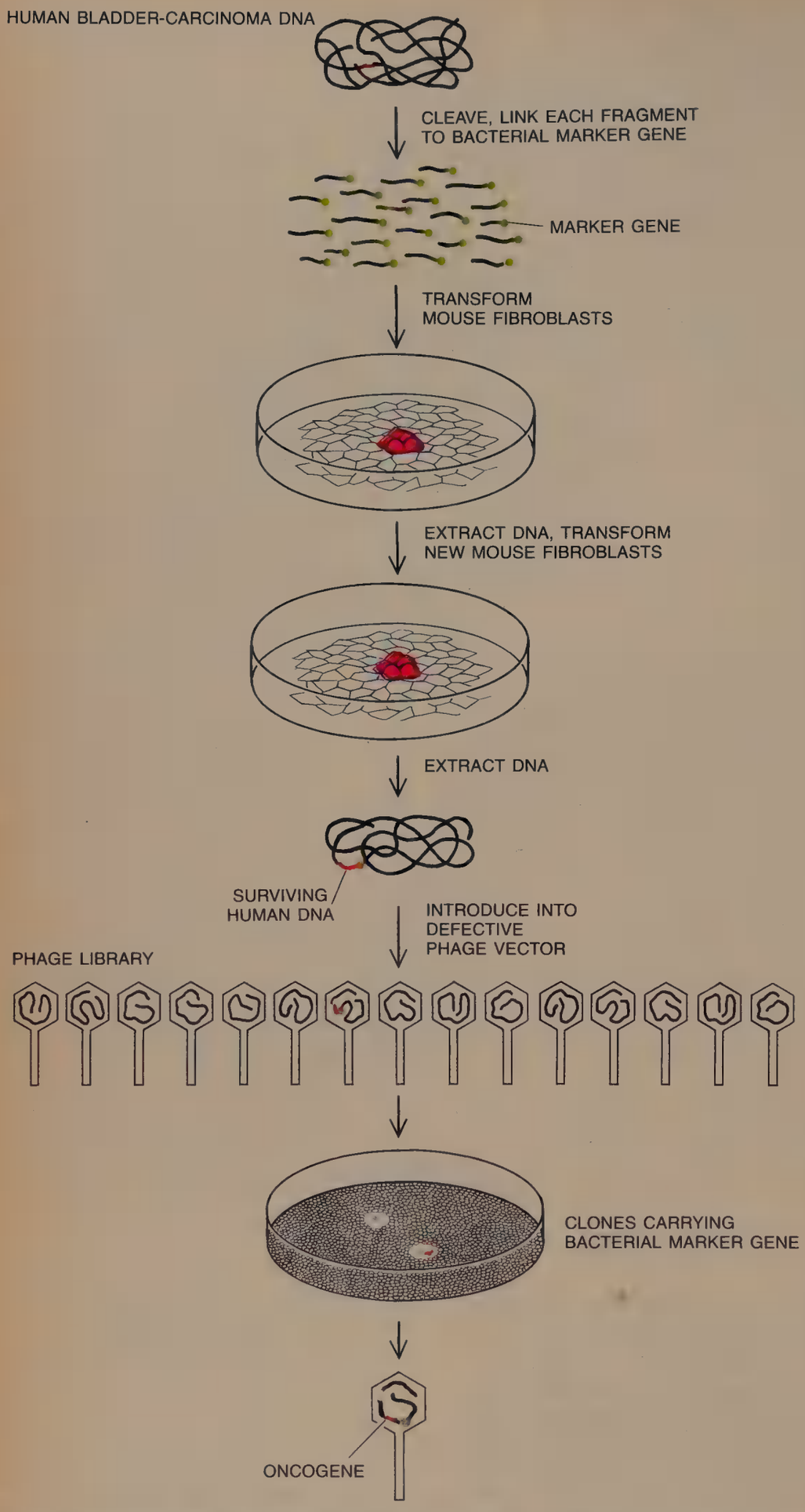
Address_____

City_____ State_____ Zip_____

Or call toll-free: **1-800-33M-HEAR** or 1-800-336-4327.

©1983, 3M

3M



BACTERIAL MARKER GENE served as a probe for Michael Wigler's group. They cleaved the bladder-cancer DNA and linked each fragment to a copy of the bacterial gene (*green*). By serial gene transfers they prepared mouse DNA carrying no human DNA except for the transforming gene (*red*), which was closely linked to a copy of the bacterial marker gene. A phage library was constructed by introducing the fragmented mouse DNA into a strain of phage that has a defect inhibiting its growth. The bacterial gene had been chosen to overcome that defect. As a result only phage clones carrying the bacterial gene grew when the library was plated on *E. coli*. Some of these clones were found to carry the closely linked bladder-cancer DNA.

clear. The proto-oncogenes would not have been conserved in the genome unless they have some vital role in normal cellular metabolism. And they have indeed been conserved. Relatives of human proto-oncogenes have been found in the DNA of a number of mammals, in chickens and even in the fruit fly *Drosophila*. The ancestors of human proto-oncogenes must therefore already have evolved when a common ancestor of human beings and flies lived, more than 600 million years ago. Proto-oncogenes would not have been kept almost unchanged over such a long time unless they were and continue to be indispensable. The precise role of such genes in normal metabolism is still not known, but there are some indications that it has to do with the control of cellular proliferation. I shall have more to say about that below.

Viral Oncogenes

The gene-transfer experiments revealed one class of oncogenes and proto-oncogenes. Another class had been discovered earlier through work on retroviruses, a group of viruses that cause cancer in some animals and whose genetic material is not DNA but RNA, which is reverse transcribed into DNA when the virus infects a cell. The retroviruses were shown to carry a single gene that is responsible for the induction of cancer: an oncogene. J. Michael Bishop and Harold E. Varmus and their colleagues at the University of California at the San Francisco School of Medicine showed that these viral oncogenes are not truly viral at all but are of cellular origin. They are proto-oncogenes that were picked up and carried (transduced) by retroviruses infecting animal cells. Somehow they are activated to become oncogenes after incorporation into the retrovirus genome and are then exploited to transform cells the viruses happen to infect. To date 17 such cellular proto-oncogenes have been defined, each of them known from its association, as an active oncogene, with a particular retrovirus.

There seemed, then, to be two sets of normal proto-oncogenes and their activated oncogenic derivatives, each set defined by a different method of activation. One group, discovered by gene transfer, comprised oncogenes that arise from the mutation of proto-oncogenes. The other, discovered previously by the virologists, comprised genes that are somehow activated by retroviruses. We now know that the two groups are not mutually exclusive. In the past year or so hybridization experiments have revealed that some oncogenes of human tumors are closely related to the oncogenes carried by several retroviruses that infect rats. The human bladder-carcinoma oncogene referred to above, for

example, is a close relative of the oncogene carried by the Harvey sarcoma virus, which acquired it from the rat genome. This means the same proto-oncogene can be activated by two independent mechanisms: by mutation in human beings and by retrovirus acquisition from rats. A second cellular proto-oncogene has been found to follow this pattern. It is activated either by incorporation into the Kirsten rat-sarcoma virus or by mutations that lead to oncogenes present in a number of human carcinomas of the colon, lung, bladder and pancreas, as well as in some human sarcomas.

The two cellular proto-oncogenes, associated respectively with the Harvey and the Kirsten viruses, are themselves related evolutionarily. They are members of a gene family, designated *ras,* that has been studied in detail by Edward M. Scolnick and his colleagues, formerly of the National Cancer Institute and now at the Merck Laboratories in West Point, Pa. A third human oncogene has recently been associated with the *ras* gene family by Wigler's group (but it has not yet been found in a retrovirus). This oncogene, called *N-ras,* is in the DNA of various leukemias, a lymphoma, a neuroblastoma, a colon carcinoma and several sarcomas. The presence of the *ras* oncogenes in a variety of tumor DNA's shows that the activation of a particular cellular proto-oncogene is not confined specifically to a single tissue. The same gene can be activated in any one of several tissues, giving rise in each to a different type of tumor.

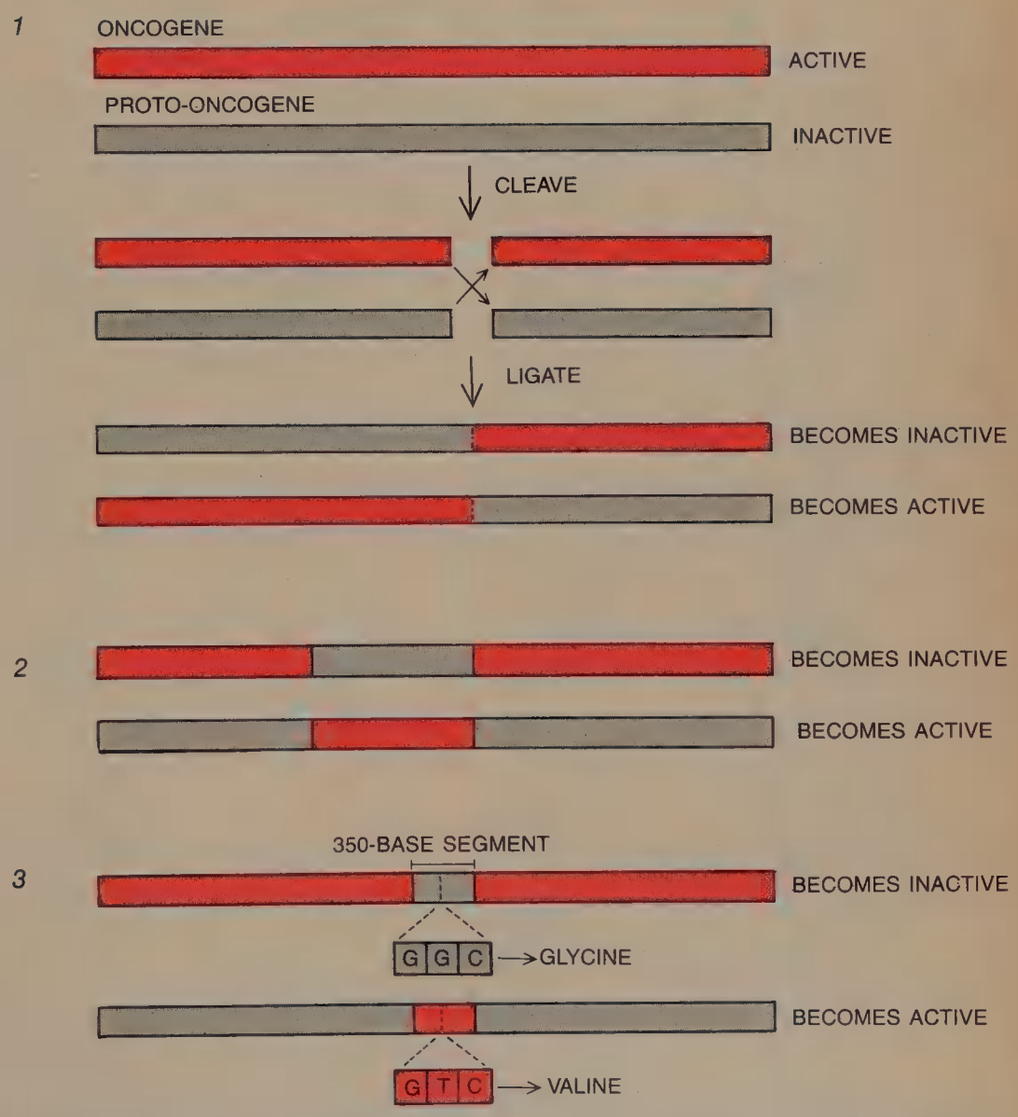
Such a simple insight represents a major advance. From the perspective of a molecular biologist cancer is no longer more than 100 diseases, each of them characterized by a different tumor type. Instead it begins to look as if there are only a small number of molecular mechanisms, which are common to all tumor types. This notion needs to be strengthened by more investigation because so far only a small number of human tumors have been studied in detail.

Defining the Mutation

A mutation was implicated as the mechanism that converts a benign proto-oncogene into an active oncogene, but the nature of the mutation remained obscure. Once cloned versions of both genes became available, the mutation could be localized with precision. In principle one could determine the entire nucleotide sequence of each of the 5,000-base-pair genes and compare the two sequences. In practice it seemed better first to define a region within the genes carrying the critical difference in sequence. This could be done by generating recombinants between parts of the two genes, applying the techniques of genetic engineering to mimic the double crossovers of classical genetics [*see illustration on this page*]. Such experiments were undertaken by three groups of investigators, in my own laboratory, in Barbacid's and in Wigler's.

First the two genes were cleaved into segments at analogous sites with restriction endonucleases, enzymes that cut DNA molecules at points defined by particular nucleotide sequences. Specific segments cut from the two genes were then joined with the enzyme DNA ligase to make two hybrid genes, part of which came from the oncogene and part from the proto-oncogene. The hybrid molecules were tested by transfer into mouse fibroblasts to see which one had lost oncogenic activity and which one had gained it.

In my laboratory Clifford J. Tabin and others narrowed down the region of interest after repeated crossovers of successively smaller segments of the bladder oncogene and its related proto-oncogene. Ultimately they found that introducing an oncogene segment only 350 nucleotides long into the proto-oncogene turned the proto-oncogene into an actively transforming molecule. This meant that the critical lesion distinguishing the two closely related genes must lie within the 350-nucleotide segment. Sequence analysis of the active and inactive versions of the segment by Ravi Dhar of the National Cancer Institute led to a dramatic and unexpected finding. The critical 350-nucleotide sequence in the oncogene and the one in the corresponding segment of the proto-oncogene differ in only one base: a guanine in the proto-oncogene is replaced in the oncogene by a thymine. This meant that a single nucleotide change, a "point mutation," in a 5,000-nucleotide normal human gene could convert it into an oncogene. For the first time a genetic lesion that can lead to the growth of a human tumor had been precisely defined. It may be that a critical event in the initiation of many cancers is the mutation of a



POINT MUTATION that converts a proto-oncogene into the active oncogene of the *EJ* bladder-carcinoma cell line was identified by first locating the segment that, in each of the genes, is the site of the critical difference and then sequencing the two segments. The proto-oncogene and the oncogene (*1*) were cleaved at the same site and the resulting segments were ligated as is shown by the broken lines. The recombined genes were tested by gene transfer to see which had become active and which had lost activity. Successively smaller segments were similarly interchanged (*2*) until a sequence only 350 bases in length was shown to be critical (*3*). There is a single mutation in that segment: a guanine (*G*) in the proto-oncogene is converted into thymine (*T*) in the oncogene, and so one triplet codon specifies the amino acid valine rather than glycine.

proto-oncogene as the result of some environmental insult to its DNA.

Recently other mechanisms of proto-oncogene activation have been described. One mechanism, chromosomal rearrangement, was first discovered by Grace L. C. Shen-Ong and Michael D. Cole of the St. Louis University Medical Center and by Rebecca A. Taub and her colleagues in the laboratory of Philip Leder of the Harvard Medical School. This mechanism relates to the activation of a proto-oncogene called *myc,* whose actively oncogenic version has been found in mouse myelomas and in a human cancer, Burkitt's lymphoma. The cellular proto-oncogene becomes separated from its usual position at the end of one chromosome and is transposed to the end of a second chromosome. There it becomes juxtaposed with genes responsible for the synthesis of immunoglobulins (antibodies). Such genes are transcribed at a high rate in the course of an immune response to an antigen. The association of the proto-oncogene with an immunoglobulin gene apparently deregulates the proto-oncogene and thereby imparts to it a novel biological function.

In other cases a proto-oncogene has been found to be present in large numbers of copies per cell instead of the normal two copies characteristic of most cellular genes. This amplification of the gene, and the resulting high level of gene expression, seems also to have an oncogenic effect. There are, in other words, a number of molecular mechanisms that can force a normal gene into the role of an oncogene. In the case of mammalian tumor genes they include point mutation, chromosomal rearrangement and gene amplification. In the case of retrovirus-associated genes the transduced gene may undergo mutation or may end up adjacent to a viral regulator that increases its level of expression.

Cancer Proteins

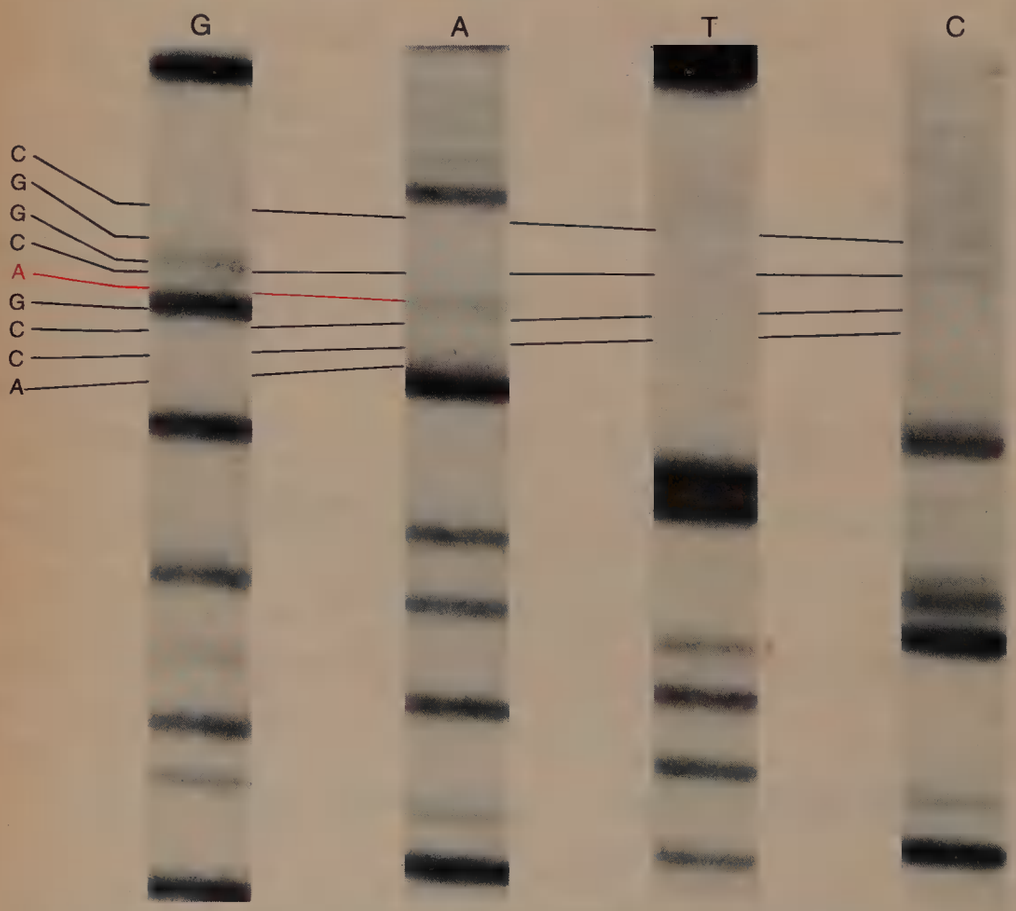
The search for the underlying mechanisms of cancer has progressed a long way, from vaguely described genetic factors to molecules of DNA, and then on to particular genes within the DNA. The search seemed to end with the finding of a precisely defined molecular alteration in one such gene. In fact, the most difficult problems are yet to be resolved.

Having granted that the creation of an oncogene may give rise to a cancer, one still has no insight into how the altered gene contributes to this process. What is the function of an oncogene? Like other genes, it encodes the structure of a protein. Scolnick's group has shown that many of the *ras* genes direct the synthesis of a single protein having a molecular weight of 21,000. The point mutation observed in the bladder-carcinoma oncogene alters the structure of the protein. Each amino acid of a protein is specified by a three-base codon in the gene encoding the protein's structure. The observed point mutation changed a *GGC* codon of the proto-oncogene into *GTC* in the oncogene. This resulted in turn in a change in the structure of the oncogene protein, which carries the amino acid valine at one site instead of the usual glycine. Apparently that single substitution causes the protein to assume novel functions enabling it to profoundly alter cellular metabolism.

Cellular elements such as these oncogene proteins must lie at the heart of the cancer process, but just how such proteins work remains unclear. One assumes that they somehow regulate the growth of cells, with the protein encoded by a proto-oncogene controlling normal growth whereas the version encoded by the oncogene forces cancerous growth.

An understanding of the precise functions of oncogene proteins may eventually make it possible to develop antagonists that inhibit those functions, and so perhaps to develop a therapy striking at one central defect of the cancer cell. There are probably, however, a number of defects. Carcinogenesis appears to be a multistep process. Evidence from a variety of sources indicates that a normal cell must suffer several independent alterations before it becomes a bona fide tumor cell. A point mutation that gives rise to an aberrant version of a single protein represents only a single step affecting a single gene. Presumably the creation of an oncogene fulfills one requirement for making a tumor, but there are other necessary steps; the oncogene may be necessary, but it is hardly sufficient. Clues are beginning to emerge as to the nature of the other steps.

A variety of tumors have been found in which two distinct oncogenes have arisen as a consequence of separate activation events. For example, William S. Hayward of the Rockefeller University and Susan M. Astrin of the Institute for Cancer Research have found that the *myc* proto-oncogene is activated in chicken lymphomas, whereas Cooper's group has found that a second proto-oncogene, designated *B-lym,* is activated in the same tumor cells. In both a Burkitt's lymphoma and a human promyelocytic leukemia two oncogenes have been found in the activated state. Such data suggest that the multistep nature of carcinogenesis may derive in part from a requirement for the activation of several



BASE SEQUENCE of the 350-base segment in the proto-oncogene and in the oncogene was determined by Ravi Dhar. This autoradiograph of an electrophoresis gel shows the critical mutation in the oncogene. The ladderlike array of bands represents successively shorter (from top to bottom) fragments of the segment. Four sets of fragments were prepared. Each set, labeled by the dideoxy method to end at occurrences of one of the four nucleotides, was run in a different channel on the gel. Careful inspection of the position of each band allows one to read off the sequence. Because this gel was prepared from the oncogene's complementary strand rather than the coding strand, the mutation appears (*color*) as adenine (*A*) rather than thymine (*T*).

WHILE SOME CARS MIGHT MAKE IT TO THE NEXT DECADE, STATISTICS SHOW VOLVOS COULD MAKE IT TO THE NEXT CENTURY.

Today people are trying to hold on to their cars longer than ever.

And while statistics show the average life expectancy of today's cars is eleven years, that figure pales in comparison to the life expectancy of today's Volvo.

Because statistics show the average life expectancy of a Volvo is over sixteen years.*

Which could mean over sixteen years of comfortable driving in seats equipped with adjustable lumbar supports that relieve tension and road fatigue.

Years of effortlessly maneuvering through parking lots due to power assisted rack and pinion steering that cuts one of the smallest turning circles of any car.

Years of clean air made possible by a Fresh Air Ventilation System that exhausts stale air from the passenger compartment, and helps prevent odors and fumes from ever getting inside.

If all this sounds like a sensible way to transport yourself to the next century, buy a Volvo.

Who knows?

By the year 2000 we may have a car that'll get you through that century, as well.

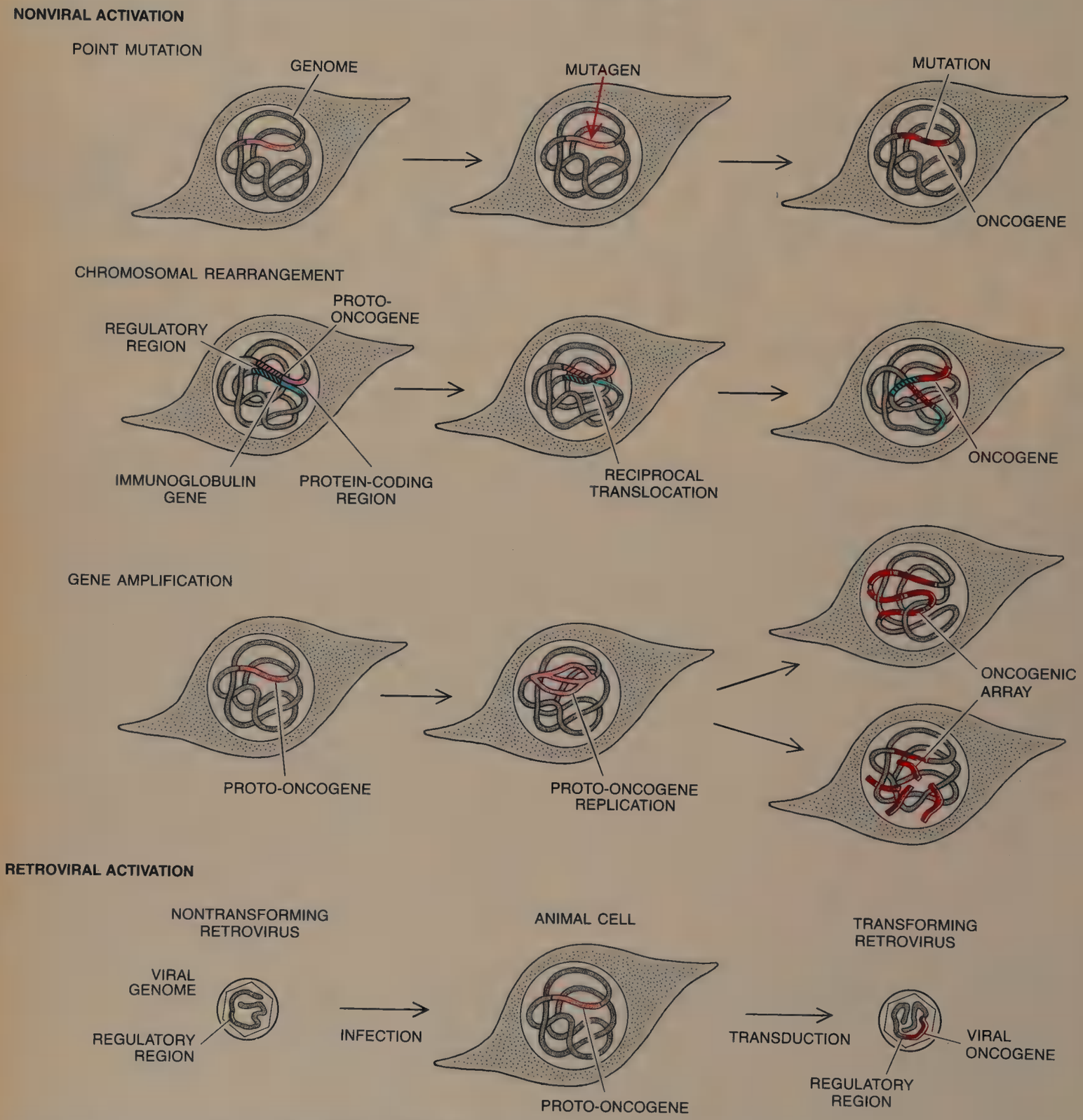
VOLVO
A car you can believe in.

*Based upon an actuarial analysis of 1981-1982 U.S. Registration Data conducted by Ken Warwick & Associates, Inc. Due to many factors including maintenance, driving conditions and habits, your Volvo may not last as long. Then again, it may last longer. Summary available at your Volvo dealer. © 1983 Volvo of America Corporation.



distinct oncogenes. Each oncogene may elicit a different alteration within the cell, and together they may collaborate to achieve important aspects of the tumor phenotype.

I have raised only two of the many issues that remain to be explored. What is most heartening is that the confluence of evidence from a number of lines of research is beginning to make sense of a disease that only five years ago seemed incomprehensible. The recent findings at the level of the gene are consistent with earlier insights into carcinogenesis based on epidemiological data and on laboratory studies of transformation. Powerful tools are available for making even more rapid advances in the coming years. The molecular mechanisms underlying cancer should be well defined by the end of this decade.



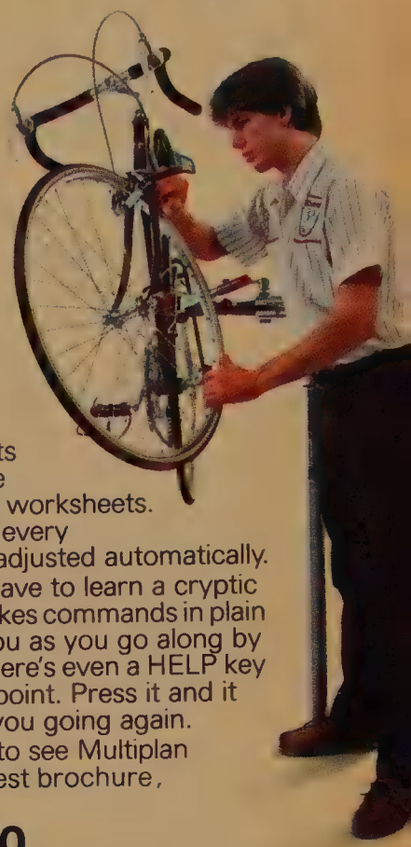
PROTO-ONCOGENE CAN BE ACTIVATED in four ways, three of them involving the alteration of cellular DNA and one involving intervention by an infecting retrovirus. One nonviral way is point mutation. If a proto-oncogene in a normal cell undergoes mutation by radiation or a chemical carcinogen, the mutation may change the information encoding a protein. The altered protein may induce one or more of the changes associated with cancerous growth. Alternatively the protein-encoding region of a proto-oncogene may become recombined, in the course of chromosomal rearrangement, with the regulatory region of an immunoglobulin gene, causing the proto-oncogene's genetic information to be expressed at an inappropriate level. A third method of activation is amplification: a proto-oncogene is somehow replicated until it is present in a cell in multiple copies, either sequentially repeated segments along the chromosome or dissociated extrachromosomal particles; again the result is inappropriate overexpression. Retroviruses, which cause cancer in many animals, acquire their oncogenes by infecting an animal cell. In the course of infection the virus picks up a cellular sequence that includes the proto-oncogene and integrates it into its own genome. The proto-oncogene may then become activated by association with a regulatory region of the viral genome, as is shown here, or by undergoing mutation while it is being carried (transduced) by the retrovirus.

"Microsoft® Multiplan™ allows me to explore more alternatives in less time. I think it leads to better management decisions."

Edward R. Schwinn, Jr.
President, Schwinn Bicycle Company

"Multiplan software helps me make better business decisions. It's as simple as that. With Multiplan on my microcomputer, I can explore a number of alternatives faster. That's really important to me in planning and tracking my totally hand-crafted Schwinn Paramount Bicycle. As with anything handmade, cost control and product planning are major considerations. For instance, Multiplan lets me study the effect of the learning curve on overall productivity. And although we have a mainframe computer, I find that Multiplan is a tool that allows me to personally examine my business options—without spinning my wheels."





A tool for decision makers. Microsoft Multiplan is a microcomputer software program that can help you too. In many ways. You can analyze cash flow. Plan budgets. Forecast income. Manage production. Multiplan allows you to set up an electronic worksheet for whatever your need may be. And, it lets you explore alternatives faster because when you make one change you immediately see the results of that change throughout the worksheet and on any related worksheets. Just change one number and every number that depends on it is adjusted automatically.

In plain English. You won't have to learn a cryptic language to use Multiplan. It takes commands in plain English. And, it will prompt you as you go along by telling you what to do next. There's even a HELP key in case you need help at any point. Press it and it gives you information to get you going again.

See it in action. If you'd like to see Multiplan in action, and receive our latest brochure, call us toll-free:

1-800-426-9400
In Washington State call 206-828-8080

We'll gladly locate the Microsoft dealers nearest you. They'll demonstrate Microsoft Multiplan and explain its compatibility with most popular personal computers. Including Apple® and the IBM® PC. It's compatible, in fact, with any computers that use the CP/M®-80 or MS™-DOS operating systems. Microsoft Multiplan is a good management tool for building bicycles. It's also a good management tool for building businesses. Like yours.

BETTER TOOLS FOR MICROCOMPUTERS

MICROSOFT™

Microsoft is a registered trademark, and Multiplan, MS, and the Microsoft logo are trademarks of Microsoft Corporation.



Solid-State Superlattices

They are crystals grown by depositing semiconductors in layers whose thickness is measured in atoms. Their properties promise to make them important in solid-state physics and in technology

by Gottfried H. Döhler

"You see things; and you say, 'Why?' But I dream things that never were; and I say, 'Why not?'" These sentences by George Bernard Shaw reflect the spirit in which a few solid-state physicists began a decade ago to speculate about the physical properties, and even the usefulness, of nonexistent semiconductors. The physicists were realistic enough to confine their thoughts to semiconductors that could exist in principle, although nature had never made them. Specifically they envisaged solid-state superlattices: semiconductors in which two materials with different electronic properties are interleaved in thin layers by depositing sheets of two semiconducting materials in alternation or by introducing impurities into layers of a single semiconducting material.

Their hopes have been honored by twofold success. First, it has turned out that superlattices can be "grown" with precision. Second, the properties of the superlattices have confirmed the investigators' predictions. Some of the properties are exotic: a paradoxical decrease in electric current with increasing voltage, for example, or the survival of free electrons in a superlattice for periods of hours rather than billionths of a second. The prediction of such properties before superlattices even existed is indicative of the progress in solid-state physics. It also signals a tendency that will gain increasing importance: the use of the quantum theory of solids to guide the engineering of new and useful materials.

As a prelude to describing the properties of solid-state superlattices it is appropriate to review the fundamental characteristics of semiconductors. Consider a pure and perfect crystal of a semiconducting material such as silicon. At a temperature of absolute zero (zero degrees Kelvin) the material would be an insulator because its electrons (the charge carriers responsible for electrical conduction in metals) would not be free to move. The explanation for this lack of mobility is that the electrons in the semiconductor can have only certain levels of energy and that the levels form energy "bands." A "band gap," an appreciable difference in energy, separates the material's "valence band," which is completely occupied by the electrons forming the chemical bonds that hold the crystal's atoms together, from its "conduction band," which is empty. At zero degrees the band gap would be insurmountable; at greater temperatures an occasional electron acquires the energy needed for it to reach the conduction band. That is, the electron becomes mobile.

What does the "occupation of a band" really mean? Suppose N atoms are bound in a solid. Each atom, if it were isolated, would have a number of energy states available to its electrons. The tenet of quantum physics known as the Pauli exclusion principle dictates, however, that each state could accept only two electrons at most and that the spin (the intrinsic angular momentum) of the two would have to be aligned in opposite directions. In the solid these conditions are changed: each state is split into a set of N substates. In a solid of appreciable size N is at least 10^{23}; thus the substates are so numerous and so closely spaced in energy that they form a quasi-continuum. Still, each state in each atom can accept only two electrons.

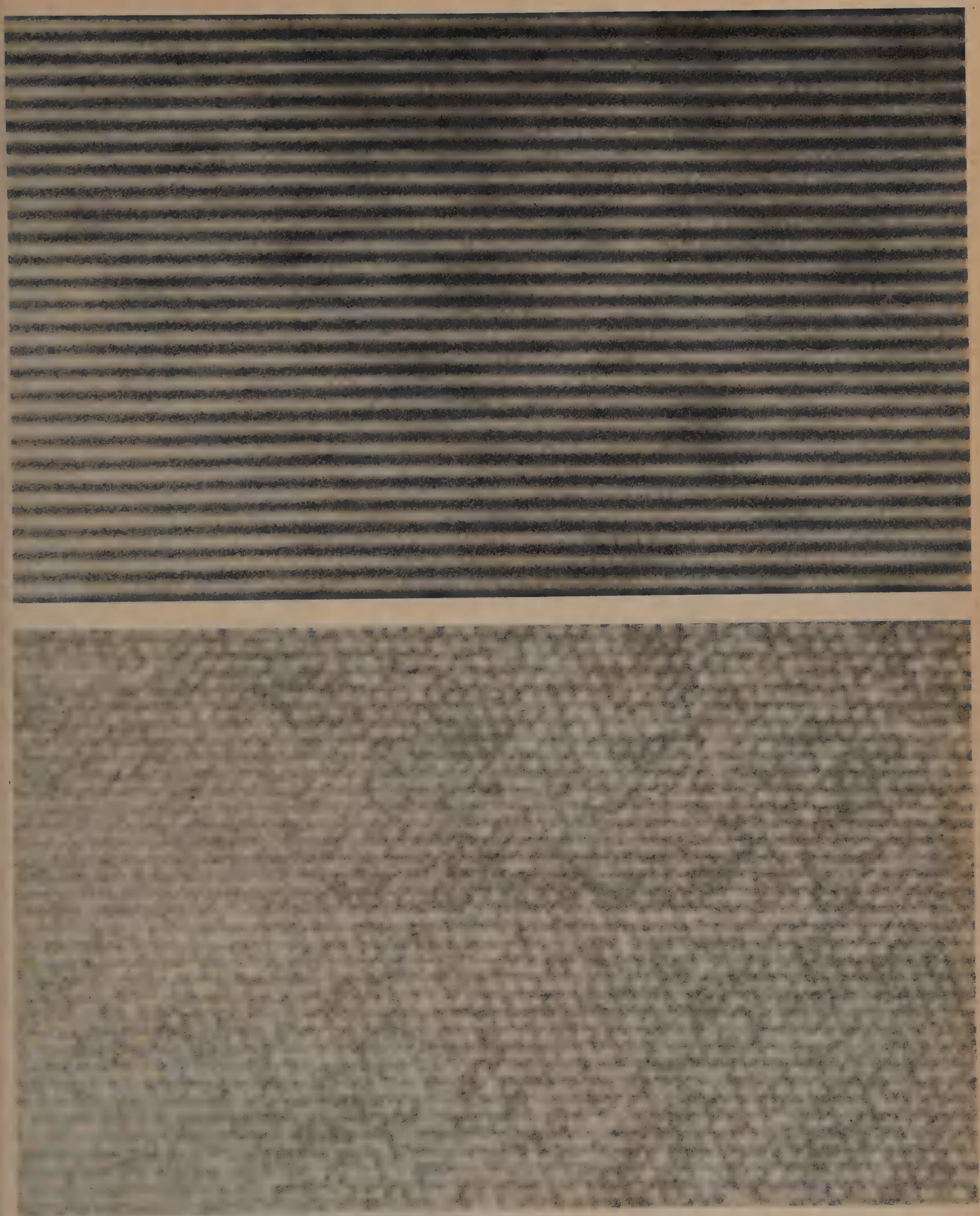
Now suppose each atom has an even number of electrons. In such atoms the states of lowest energy are fully occupied, each state by two electrons; hence the valence band is full. Suppose further that the valence band and the conduction band do not overlap in energy, yet are not very far apart. The solid is then a semiconductor. (If the valence band and the conduction band did overlap, electrons from the uppermost levels of the valence band would "flow" freely into lower-lying conduction-band levels.) Suppose the number of atoms is odd, so that the uppermost state in the valence band is only half filled. Then the solid is a simple type of metal.

It is at least plausible from this discussion why a simple metal is a conductor. The electrons in the metal are readily accelerated by an electric field because the difference in energy between an occupied substate and unoccupied substates in a quasi-continuum of 10^{23} substates is negligibly small. In contrast the electrons in a semiconductor must be transferred across the band gap and into the conduction band before they can gain energy gradually in an unoccupied quasi-continuum. The transfer can be achieved in several ways, such as when the semiconductor absorbs a photon (a quantum of light) with an energy greater than the width of the band gap.

One further characteristic of semiconductors is also fundamental. Each electron excited into the conduction band of a semiconductor leaves behind a "hole" in the valence band, that is, it leaves behind the absence of an electron in one atom or another. The absence can shift from atom to atom; hence the hole acts in every respect as a carrier of charge, and it contributes to the crystal's conductivity. (In a way, then, the terms conduction band and valence band are misleading.) The charge of an electron is negative; that of the hole is positive.

Suppose a free electron and a hole are generated in a semiconductor by the absorption of a photon whose energy exceeds the band gap. They will lose the excess energy quite fast by warming the crystal. Thus the electron will "tumble downward" to the lower edge of the conduction band. The hole, meanwhile, can be thought of as "ascending" to the upper edge of the valence band. If the two recombine, the semiconductor will emit a photon whose energy is equal to that of the band gap. Clearly the energy value of the lower edge of the conduction band and that of the upper edge of the valence band are important quantities for describing a semiconductor.

The first solid-state superlattice to become a reality was a compositional (or



COMPOSITIONAL SUPERLATTICE is a periodic array consisting of alternating layers of two different semiconductors. In the electron micrograph at the top the semiconductors are gallium arsenide and aluminum arsenide. In particular, each dark band is six atomic planes of gallium arsenide; each bright band is four atomic planes of aluminum arsenide. The distance between successive atomic planes is 2.83×10^{-10} meter. In the electron micrograph at the bottom a single interface between a layer of gallium arsenide (*top half*) and a layer of gallium aluminum arsenide (*bottom half*) is shown. The interface, which runs across the middle of the image, is a single atomic plane thick. The distance between planes is 3.26×10^{-10} meter. Each bright point in the image is a column of atoms viewed head on in the superlattice crystal. The electron micrographs were provided by Pierre M. Petroff and Arthur C. Gossard of Bell Laboratories.

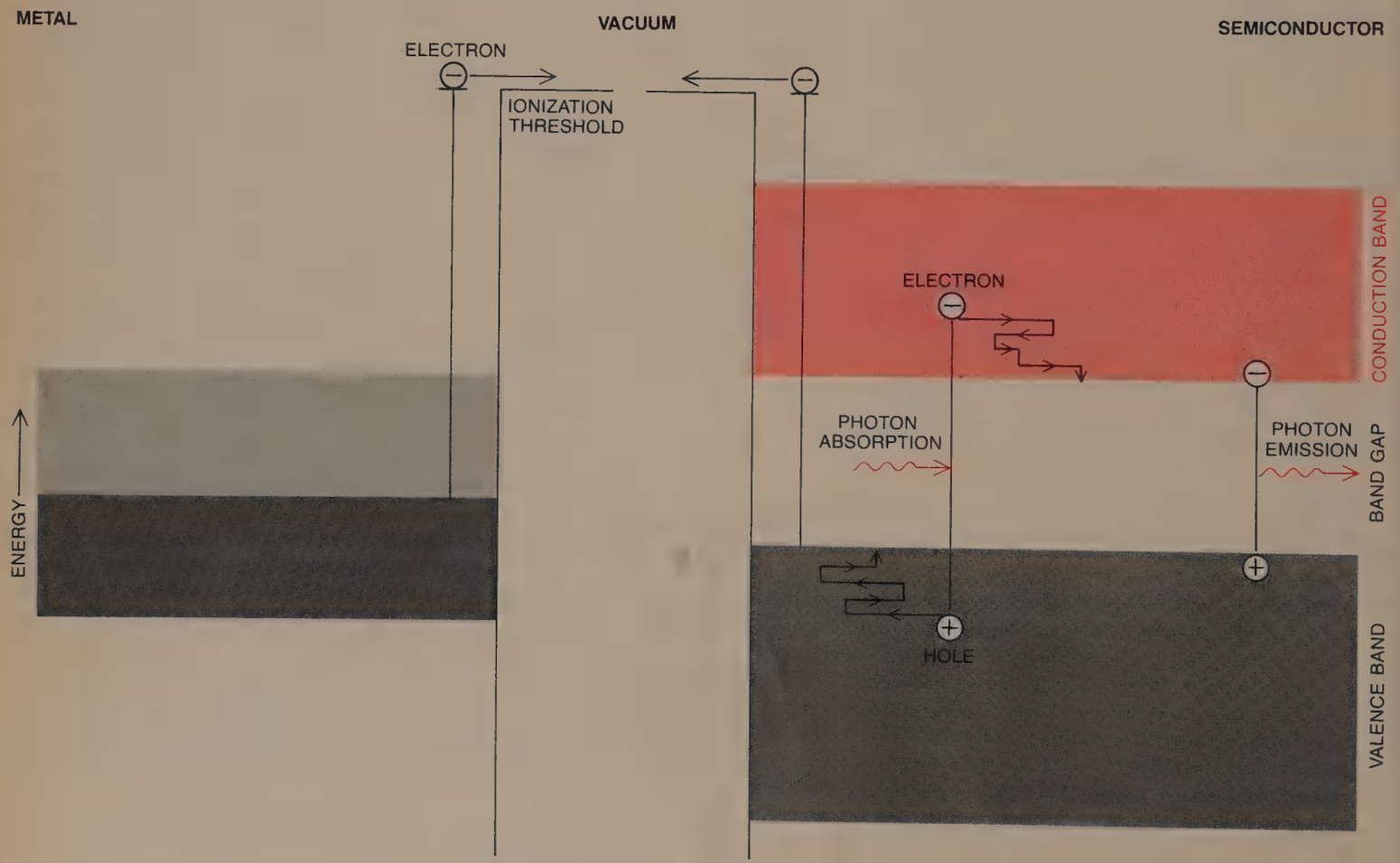
heterostructure) superlattice. This novel semiconducting material had been proposed in 1970 by Leo Esaki and Ray Tsu of the Thomas J. Watson Research Center of the International Business Machines Corporation. It was grown a few years later by Esaki and Tsu and their colleagues.

A compositional superlattice is a periodic array of ultrathin layers of two different semiconductors in alternation. Each layer is no more than a few hundred atoms thick. The semiconductors are chosen so that their band gaps are different. Thus the simple banding of energy levels in a bulk semiconductor yields to a more complex pattern resulting from the spatial variation in the material's electronic properties. The most straightforward way to determine the pattern is to treat the superlattice as simply a series of semiconducting layers, with each layer contributing its characteristic properties. More sophisticated calculations show that this strategy gives satisfactory results, even in the case of extremely thin layers. It can be said, therefore, that the periodic alternation of layers gives rise to a periodic alternation in electric potential. Specifically, each layer of the semiconductor with the smaller band gap produces what is called a potential well [*see illustration on opposite page*]. Inside each potential well only certain energy states are available to conduction-band electrons. (Again each state is split into a quasi-continuum, or miniband.) The situation is similar to that of a crystalline solid, where periodicity is established by the presence of atoms at regular intervals. Each atom creates a potential well, and in each atom only certain energy states are available to electrons.

There is an important difference, however. The electronic properties of the atom are predetermined by nature. The electronic properties of the superlattice can be designed. In the first place the values of the energy levels available to electrons can be tailored by the appropriate choice of semiconductors and of the width of their layers. (Basically the same is true of the levels available to holes.) In addition the width of each miniband can be tailored. The width (which turns out to be determined by the strength of the interaction between neighboring potential wells) increases with decreasing width of the layers of the semiconductor whose band gap is greater.

In practice the semiconductor with the smaller band gap can be a material such as gallium arsenide (GaAs) and the one with the larger band gap can be aluminum gallium arsenide (AlGaAs). (Any ratio of aluminum atoms to gallium atoms is possible, but the total number of aluminum and gallium atoms must equal the number of arsenic atoms.) Calculations for a compositional superlattice consisting of layers of such materials show that the minibands are much narrower than the bands in any bulk semiconductor and also that the minibands are separated from each other in the conduction band by relatively large "minigaps." The only requirement is that the thicknesses of the two types of layers be in the range from four to 10 nanometers. (A nanometer is a billionth of a meter.)

The minibands provide the ideal environment for the phenomenon of solid-state physics known as Bloch oscillation, which seems to be undetectable in any solid occurring in nature. It relies on the fact that a high electric field tilts the bands in a semiconductor. The slope of the tilt is e (the charge of an electron) times the voltage applied to the crystal, divided by the length of the crystal; thus the slope increases with increasing voltage. In a conventional semiconductor the electrons in the conduction band are driven toward the upper edge of the tilted band [*see upper illustration on page 148*]. Long before they reach it, how-

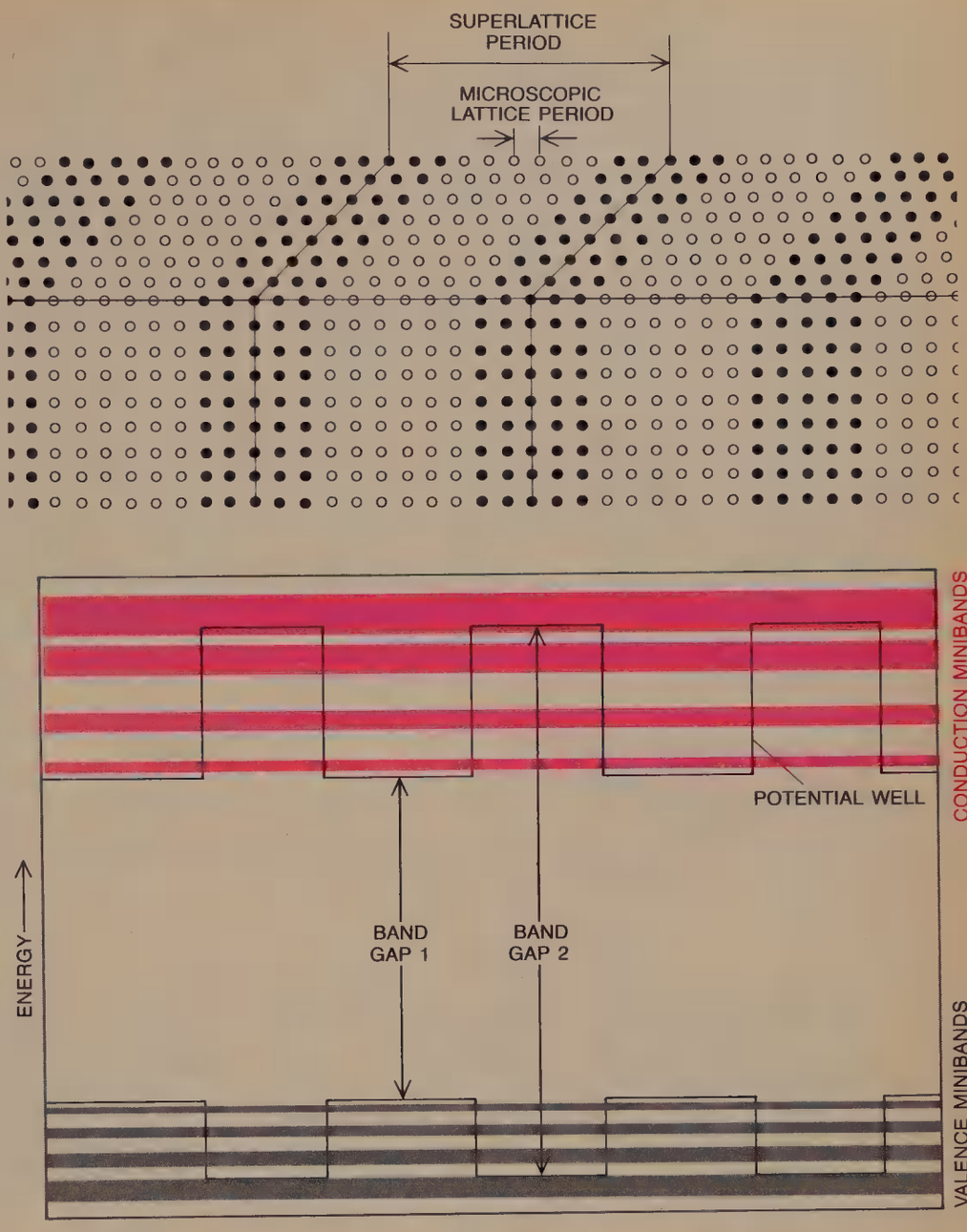


CONTRASTING ELECTRONIC PROPERTIES of a metal and a semiconductor are displayed in diagrams that show the energy levels available in each material to carriers of electric charge. In the metal (*left*) the carriers are electrons, which move freely in a partly filled band of available energies. In the semiconductor (*right*) the electrons that form bonds between atoms fill up a "valence band." They are excited into a "conduction band," which otherwise is empty, by the absorption of photons (quanta of light) whose energy exceeds the "band gap." The excitation leaves "holes" in the valence band. When electrons and holes recombine, their energy is dissipated by photons.

ever, they lose the energy they have gained from the field by emitting "phonons," that is, by exciting thermal vibrations in the crystal.

In a superlattice the situation is different. The minibands are narrow, and so the probability of an electron's reaching the upper edge of a tilted miniband is high. At the upper edge the electrons are reflected, in essence because they cannot pass through a minigap into the next miniband. After such a reflection (the quantum phenomenon known as Bragg reflection) they turn back toward the bottom edge of the miniband. Indeed, they can be reflected repeatedly so that they shuttle back and forth between the upper and the lower edge of the miniband, performing many of these Bloch oscillations before they emit a phonon and "fall" to a lower energy.

The average distance by which the emission of a phonon shifts the average position (the center of mass) of an electron along the conduction band decreases with increasing tilt of the band. Thus an increase in the voltage applied to a superlattice can have the curious effect of making the current flowing through it decrease. In other words, the crystal can exhibit a negative resistance: it can refrain from consuming energy like a resistor and instead can feed energy into an oscillating circuit. A superlattice might therefore serve as the active element of an electromagnetic wave generator. Unlike other negative-resistance devices it would react almost instantaneously to a voltage; hence it could generate microwave radiation with wavelengths of less than one millimeter. Efficient generators for this range of wavelengths are not available today.



PROPERTIES OF A COMPOSITIONAL SUPERLATTICE result from the periodic alternation of its two different semiconductors. The two are chosen to have band gaps that differ in width, and the difference creates potential wells: abrupt drops in the potential energy inside the crystal (*bottom chart*). In turn the potential wells split the valence band (*gray*) and the conduction band (*color*) into minibands. The crystal can be tailored (that is, the width and the energy of the minibands can be set) by the choice of semiconductors and of the thickness of their layers.

A remarkable thing about the Bloch oscillations in a compositional superlattice is that Esaki and Tsu predicted them before superlattices existed. Then came the fabrication of superlattices, by the technique of molecular-beam epitaxy, which Alfred Y. Cho and John R. Arthur, Jr., had developed at Bell Laboratories in the late 1960's. In this technique crystals are grown by directing beams of their constitutent atoms or molecules at a substrate such as a gallium arsenide wafer. The beams deposit matter at a rate of about one atomic layer per second. A few years later Esaki, working with L. L. Chang, reported the first experimental verification of the prediction that had started it all.

This first success was soon followed by another. Raymond Dingle, William Wiegmann, Arthur C. Gossard and Charles H. Henry of Bell Laboratories had been investigating the absorption of light by a compositional superlattice consisting of gallium arsenide and gallium aluminum arsenide. They could predict that photons would be absorbed only if their energy equaled or exceeded a threshold value determined not by the band gap of either semiconductor alone but by the effective band gap: the difference in energy between the least energetic conduction miniband and the most energetic valence miniband [*see illustration above*]. Further, they could predict that the absorption would sharply increase near photon energies corresponding to the differences in energy between other pairs of minibands.

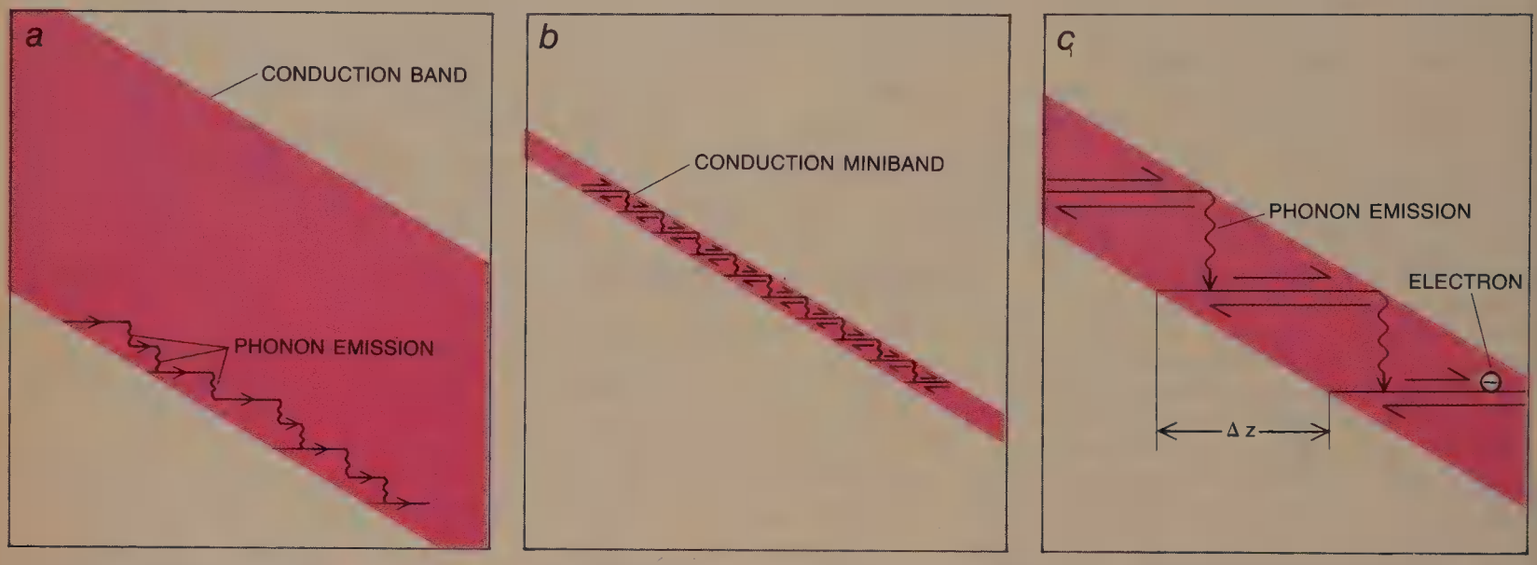
Experiments confirmed these predictions; thus quantum physics proved itself in a novel manmade material. Conversely, the confirmation of the predictions established that the superlattice had indeed been grown with precision. This raised the possibility that superlattices might serve the study of complex quantum-mechanical phenomena. Specifically, the interactions among the particles in a quantum-mechanical many-body system might be investigated under simple and well-defined conditions if a rectangular potential well engineered with great precision in a superlattice were occupied by many charge carriers.

In this respect too superlattices exhibit remarkable properties. So far I have described only semiconducting materials in which charge carriers are generated by the absorption of light. Instead of such carriers, which disappear rapidly owing to electron-hole recombination when the excitation of the crystal has stopped, one can introduce a permanent population of electrons or holes by "doping" the semiconductor with impurity atoms. Atoms whose number of valence electrons (electrons that can enter into chemical bonds) is one greater than

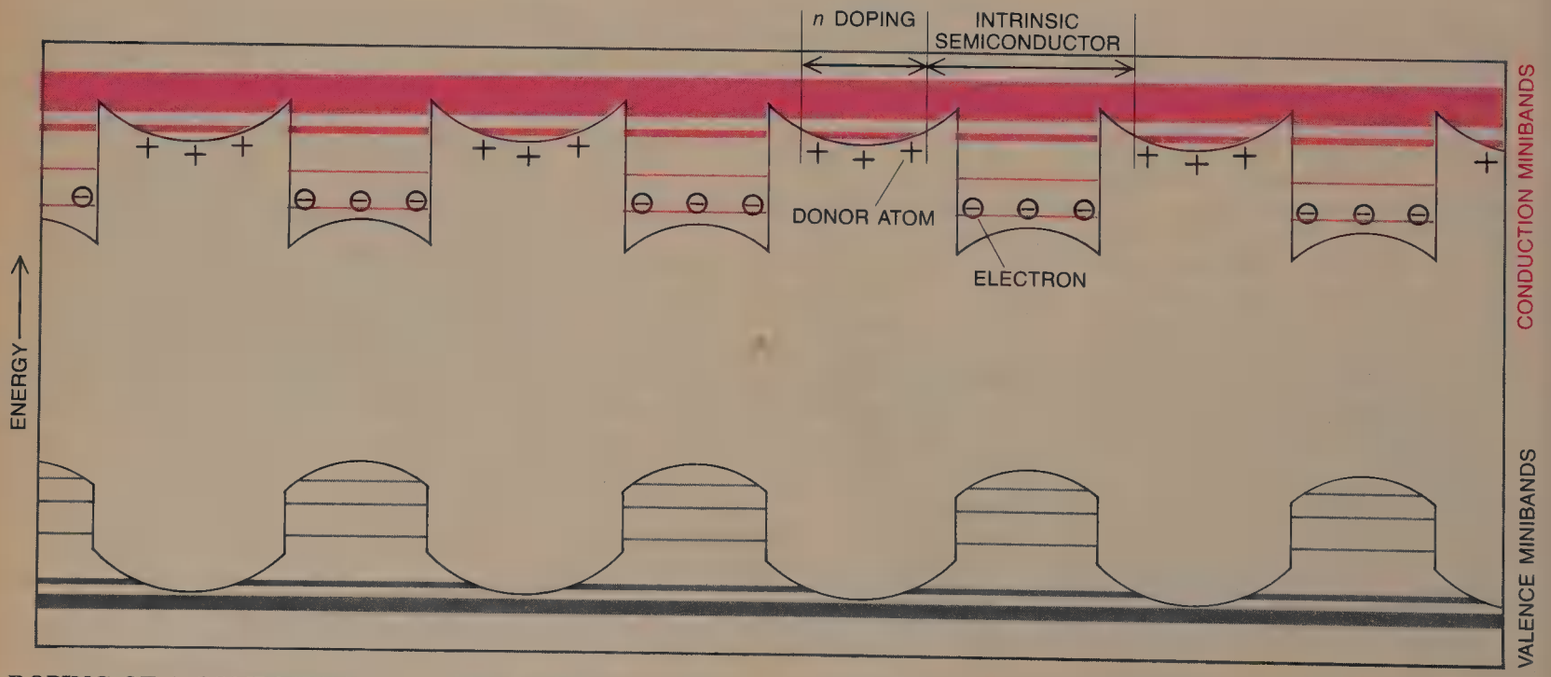
that of the atoms for which they substitute in a host crystal are donors: they can easily give away the loosely bound excess electron. Thus a semiconductor doped with donor atoms is no longer an insulator even at a temperature of absolute zero: it is an electron conductor, or *n*-type semiconductor, where *n* stands for the negative charge of the electron. Conversely, atoms whose number of valence electrons is one lower than that of the host-crystal atoms are acceptors: they can easily capture a valence electron from a host-crystal atom, thereby producing a hole. Hence a semiconductor doped with acceptor atoms is a hole conductor, or *p*-type semiconductor, where *p* stands for the effective positive charge of the hole.

Unfortunately the donors and acceptors in a semiconductor do more than provide charge carriers. At the same time they reduce the mobility of the carriers. The reason is that a donor, having given up an electron, or an acceptor, having gained one, becomes an ion and therefore gets an electrostatic field that impedes the motion of charge carriers by scattering them. For the engineer this is disadvantageous: it means the electrons and holes will not move as fast as one would wish. For the investigator of quantum-mechanical many-body systems it is also disadvantageous: it means the interaction among electrons will be obscured by the interaction between electrons and impurity atoms, which is of comparable strength.

In compositional superlattices the disadvantage can be overcome by a simple trick. A doping of donor atoms can be confined to the layers of the semicon-



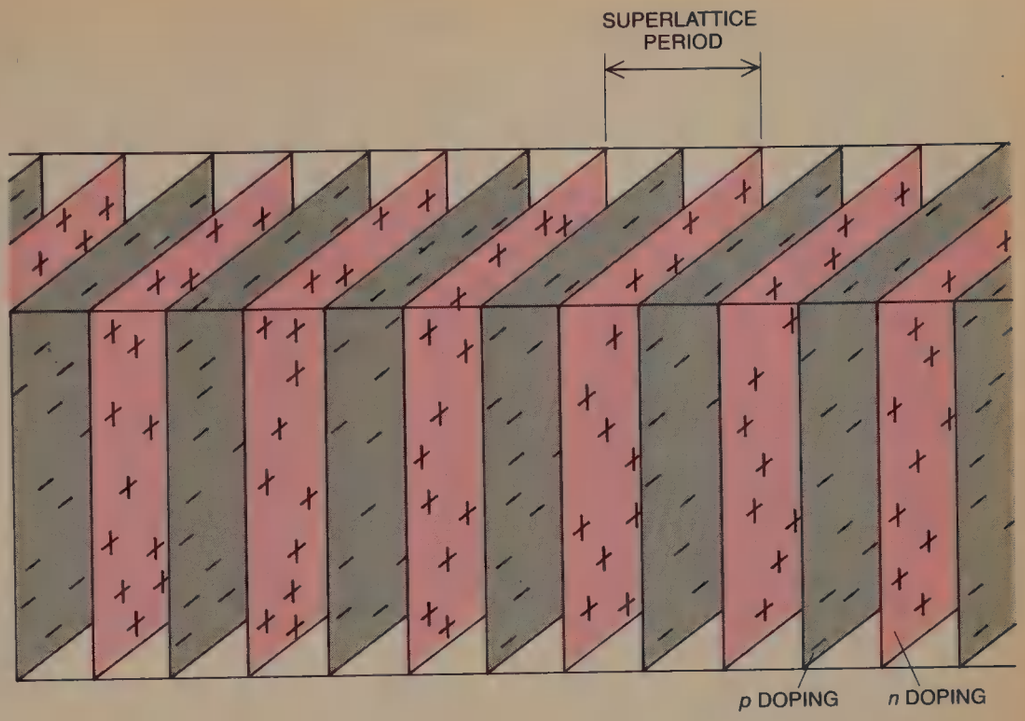
BLOCH OSCILLATION depends on the fact that a voltage accelerates electrons in the conduction band of a semiconductor but at the same time tilts the band. Electrons are therefore propelled toward the upper edge of the band. In a typical semiconductor (*a*) they never arrive. Instead they emit phonons, that is, they excite thermal vibrations in the crystal lattice and "fall" to a lower energy. In a superlattice, where the conduction band is split into narrow minibands, they do arrive. Specifically, they arrive at the upper edge of the lowest conduction miniband (*b*), and on arrival they are reflected. They may oscillate repeatedly between the miniband's edges before they emit a phonon. Since the miniband is tilted, the emission moves the electron's center of mass a certain distance Δz (*c*). That distance decreases with increasing tilt; thus an increase in voltage can have the curious effect of decreasing the current that flows through the crystal.

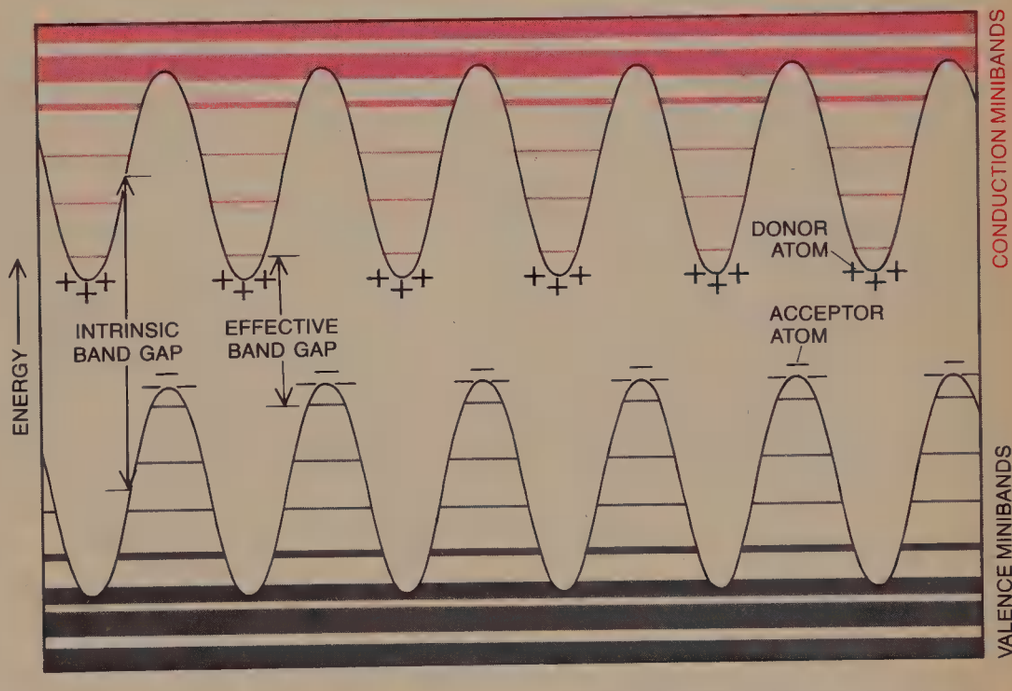


DOPING OF A COMPOSITIONAL SUPERLATTICE (the introduction of impurities into the lattice of the crystal) increases the mobility of electrons. The semiconductor with the greater band gap is doped with donor atoms, each of which contributes an electron to the crystal and thereby becomes a positive ion (*plus signs*). The electrons (*minus signs in circles*) then populate the lowest available conduction miniband, which is in the other layers. The resulting spatial separation between the charge carriers and the positive ions reduces their interaction, and so the conductivity of the crystal can be several hundred times greater than it would be in a uniform semiconductor with the same carrier concentration. The bending of the periodic potential wells in the superlattice is caused by the electrons and the ions.

ductor that has the larger band gap [*see lower illustration on opposite page*]. Each donor will contribute a free electron to the crystal, and the electrons will seek the available states that have the lowest energy. These states, however, will be in the layers of the other semiconductor (they will be the lowest conduction minibands there). The electrons will thus gain spatial separation from the donors, and the conductivity of the material may be made several hundred times greater than it would be in a uniform semiconductor with the same concentration of carriers. The trick was first tested experimentally in 1978 by Horst L. Störmer, Dingle, Gossard, Wiegmann and Ralph A. Logan at Bell Laboratories. Once again the experiment confirmed the prediction. Today improved designs and crystal-growing techniques in laboratories in the U.S., Europe and Japan seem to be yielding new records in electron mobility (and new gratification for engineers interested in fast electronic devices) every few weeks.

The more the investigations of compositional superlattices proved successful, the more I grew convinced that another kind of superlattice was also practicable. It would be a doping superlattice, that is, a single bulk semiconductor modulated only by periodic *n* and *p* doping. The problem was to capture the interest of people who might be able to grow such structures. It seemed to me that the technical difficulties facing the growers of a doping superlattice would be no worse than those that had faced the growers of compositional superlattices, and the electronic and optical properties of a doping superlattice, which I had predicted as early as 1972, were certainly no less dramatic than those of a compositional superlattice.





DOPING SUPERLATTICE is a periodic array consisting of layers of the same semiconductor doped in two different ways (*top drawing*). In the *n* layers donor atoms (*plus signs*) contribute electrons; in the *p* layers acceptor atoms (*minus signs*) bind electrons. The resulting distribution of electric charge creates a set of potential wells (*bottom chart*). The value of the effective band gap (the energy difference between the highest valence miniband and the lowest conduction miniband) depends on the choice of dopant concentrations and of layer widths.

Consider a periodic array of *n* and *p* doped layers, separated, perhaps, by undoped layers of the same semiconductor. (These latter layers are *i* layers, where *i* stands for intrinsic, and so the superlattice is an *n-i-p-i* crystal.) A certain fraction of the donors and acceptors would be ionized, because electrons leaving the donors would recombine with holes leaving the acceptors as long as the process continued to lower the total energy of the material. The resulting charge in the doped layers (positive charge in each *n* layer, negative charge in each *p* layer) would produce a periodic electrostatic potential, and this potential would modulate the conduction and valence bands much as it does in a compositional superlattice.

One great advantage of a doping superlattice is the flexibility with which it can be tailored. For one thing, any semiconductor can be the host material for a doping superlattice, provided only that both *n* and *p* doping are possible. In contrast, the choice of two semiconductors for a compositional superlattice turns out to be strongly constrained by the requirement that the interatomic distances in one of them match at least approximately the distances in the other. Moreover, the effective band gap of a doping superlattice can be given any value from zero to the band gap characteristic of the undoped host material by choosing the appropriate combination of doping concentration and layer thickness. If the doping concentration and layer thickness are sufficiently great, the superlattice becomes a semimetal: a material with a band gap of zero and a residual concentration of free electrons and holes.

Even after a doping superlattice is fabricated it can be tuned. That is, the electronic and optical properties of a doping superlattice can be designed to have certain values and then the values can be modulated within wide limits by a weak excitation of the crystal, say the absorption of a low-intensity light signal

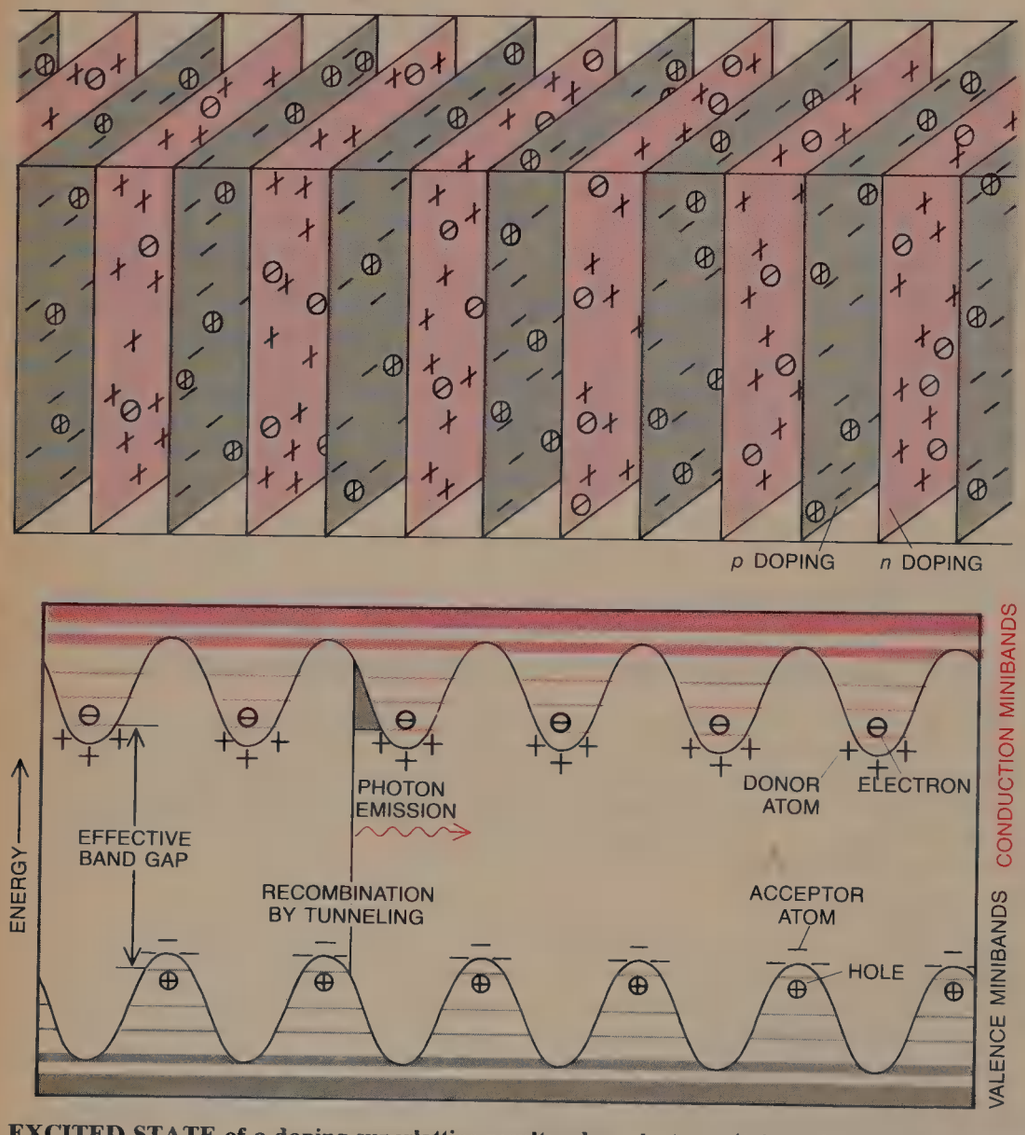
or the application of a small electric current. This tunability, which differentiates doping superlattices from all other semiconductors, arises from the spatial separation between the electrons and the holes that the excitation produces. The electrons are in the *n* layers; they "tumble downward" into the lowest conduction miniband. The holes are in the *p* layers; they "ascend" to the highest valence miniband.

The *n* and *p* layers are separate; thus the electrons and holes lack partners for recombination. If they obeyed the laws of classical physics, they would never recombine, unless, say, thermal energy in the crystal pushed them to the same location. Recombination by this mechanism is likely only at high temperature. By the laws of quantum physics, however, another means of recombination exists: the particles can "tunnel" through a potential barrier. Still, the probability of recombination by tunneling decreases dramatically with increasing height and width of the barrier. Calculations show, therefore, that doping superlattices can be tailored so that free electrons and holes can have lifetimes ranging from a few nanoseconds (the lifetime of the charge carriers in a uniform semiconductor) up to several hours.

It follows that in a doping superlattice large deviations from the ground state (the state of lowest total energy) can be made to decay very slowly. Indeed, the deviations can be stabilized by simply injecting electrons and holes (hole injection amounts to electron extraction from the valence band) or by causing their generation at a low rate. The concentrations of electrons and holes in a doping superlattice are therefore tunable quantities. The tunability is crucial. In the *n* layers the electron concentration partly neutralizes the positive charge of the ionized donor atoms. In the *p* layers the hole concentration partly neutralizes the negative charge of the ionized acceptor atoms. Hence the superlattice potential flattens, and the effective band gap increases.

Perhaps it will not surprise the reader to learn that the electrical conductivity, the optical absorption, the light emission, in fact essentially everything about the electronic and optical properties of a doping superlattice, can be tuned. The electron-hole pairs created by the optical excitation of the crystal become separated with great efficiency. Since recombination is unlikely, their concentrations rise; thus the potential flattens and the effective band gap widens. The flattening increases the likelihood of tunneling; thus the lifetime of the charge carriers decreases. Finally the rate of recombination becomes equal to the rate at which electron-hole pairs arise. The crystal is now in a steady state with its properties tuned to specific values.

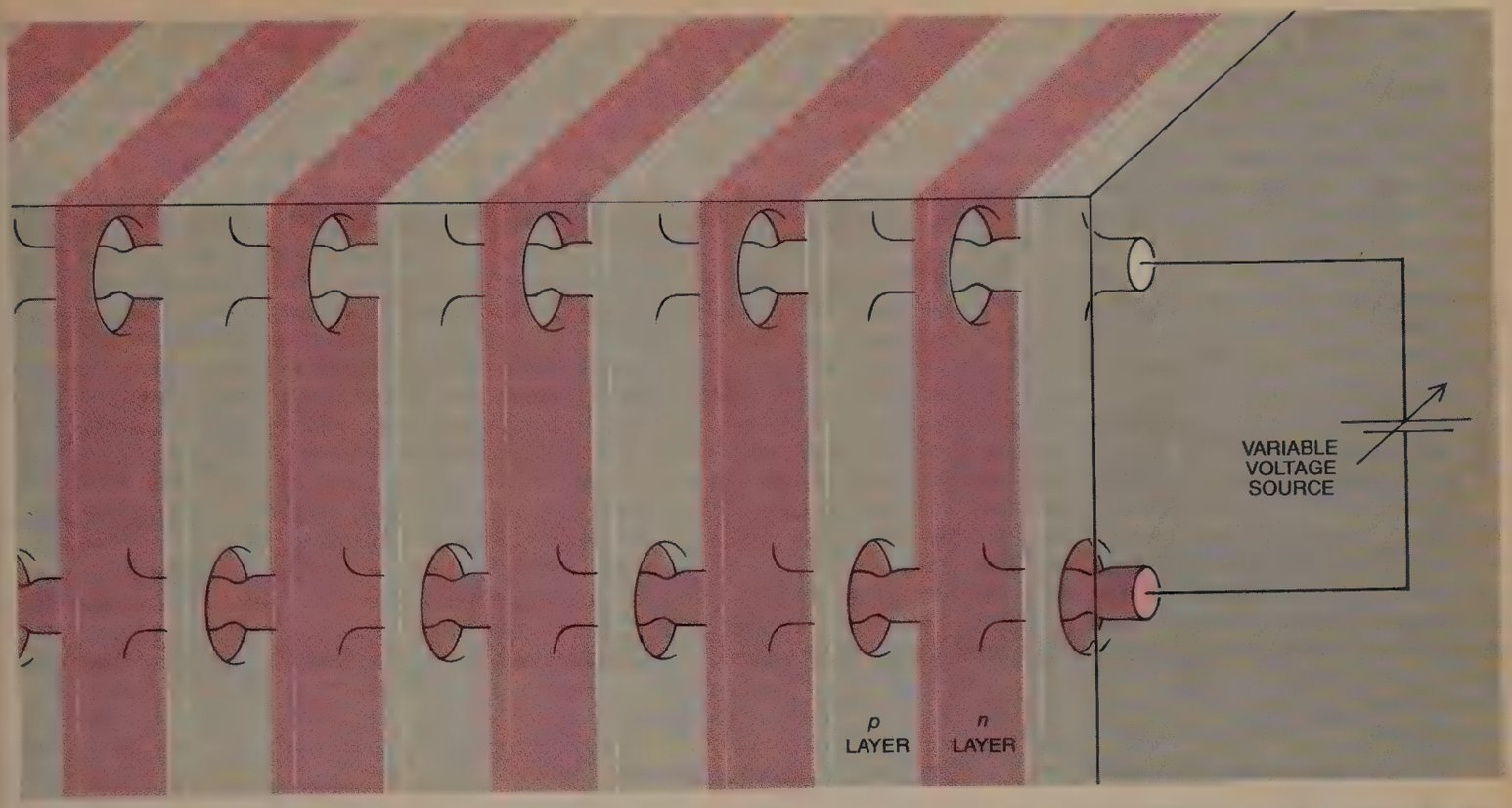
At first the excitation of the crystal by the injection of electrons into only the *n* layers and holes into only the *p* layers may seem to be an insoluble problem in microengineering. Actually it can be managed rather easily. A selective electrode to the *n* layers or the *p* layers can be a region of strong *n* or *p* doping respectively that extends through the superlattice [*see illustration on opposite page*]. At the intersection of such an electrode with a layer of the opposite type of doping the charge carriers one does not want to enter the layer in fact are kept out by electric fields. In the *n* layers the ionized donors repel the holes that would otherwise enter; in the *p* layers the ionized acceptors repel the electrons.



EXCITED STATE of a doping superlattice results when electrons (*minus signs in circles*) enter the *n* layers and holes (*plus signs in circles*) enter the *p* layers. These charge carriers partially neutralize the charge of the ionized dopant atoms, so that the potential wells flatten and the effective band gap increases. Thus the properties of a doping superlattice can be tuned by current injection or by the optical excitation of electron-hole pairs. The electrons and holes recombine if electrons "tunnel" through a potential barrier (*gray*). The rate of tunneling depends on the width and height of the barrier; hence the lifetime of charge carriers can also be tuned.

The investigator who finally undertook to grow a doping superlattice was Klaus Ploog of the Max Planck Institute for Solid-State Physics in Stuttgart. Thanks to his care and that of his colleagues, the effort, which was made in 1980, succeeded immediately. The host material was gallium arsenide; the dopants were silicon atoms, serving as donors, and beryllium atoms, serving as acceptors. The method of growth was molecular-beam epitaxy.

First the predicted relation between the concentration of charge carriers and the effective band gap was tested by measuring the electron and hole conductivity in the layers of the lattice. H. Künzel, working with Ploog, J. Knecht, A. Fischer and me, found that the conductivity vanished at values of electric potential for which the concentration of carriers should be zero. Next the predicted optical properties were tested. In one experiment, done by Künzel, Ploog and me, the absorption of light by a doping superlattice generated charge carriers with lifetimes of more than 1,000 seconds. This represents a trillionfold improvement over the lifetimes of



SELECTIVE ELECTRODES to the *n* and *p* layers of a doping superlattice are columns of material strongly doped with *n*- and *p*-type impurities respectively. Ionized donors in each *n* layer repel holes in the *p* electrode; ionized acceptors in each *p* layer repel electrons in the *n* electrode. As a result a voltage difference applied across the electrodes sends holes exclusively into *p* layers and electrons exclusively into *n* layers, thereby tuning the crystal. The injection of charge carriers stops when the effective band gap equals the voltage difference.

charge carriers in typical semiconductors. A similar factor of improvement turned out to apply to the superlattice's sensitivity to light. Clearly, doping superlattices are worth examining for their usefulness in light detectors.

The testing of optical properties also enabled us to verify the predicted tunability of doping superlattices. The most impressive proof was provided by experiments with the luminescence of the crystals. Here the investigators I have named, working with H. Jung, D. Olego, W. Rehm and H. J. Stolz, measured the spectrum of the light emitted by a doping superlattice under various intensities of the illumination that served as excitation. At increasing intensities of the incident light, the carrier concentration, the effective band gap and consequently the energy of the photons the superlattice emitted proved to increase by as much as 20 percent. Similar increases have since been observed in doping superlattices excited by the injection of charge carriers. Hence it appears doping superlattices are also worth examining for their usefulness as modulators of light, say in electronic systems that process laser beams.

Of course, the emission of light by a superlattice requires that electrons and holes recombine, which may seem inconsistent with charge-carrier lifetimes on the order of 1,000 seconds. Actually it is not inconsistent. A doping superlattice may indeed exhibit charge-carrier lifetimes in the 1,000-second range when it is near its ground state and the effective band gap is close to zero. (At a band gap of zero recombination can no longer decrease the total energy of the crystal; thus the carrier lifetimes are long.) Under strong illumination, however, the effective band gap approaches the band gap characteristic of the undoped semiconductor. The barrier to tunneling flattens out and the charge-carrier lifetimes shorten.

If one seeks, then, to produce a doping superlattice with high luminescence at photon energies far less than the band gap characteristic of the undoped semiconductor, one will tailor it in the appropriate way. Specifically, one will give it narrow layers and a high degree of doping. The narrow layers entail narrow potential wells, and they will ensure a degree of "transparency" for free electrons in the crystal even if the height of each potential well is great.

The list of the properties of doping superlattices predicted by quantum physics and then verified by experiment must now include the experimental verification of the predicted energy differences between minibands. Investigators at the Max Planck Institute in Stuttgart, including G. Abstreiter and P. Ruden, directed photons into a doping superlattice made from gallium arsenide. They measured the energy of the photons that bounced back out. Some of the scattered photons had excited electrons from populated minibands into higher, empty minibands and so had lost that amount of energy.

Today solid-state superlattices represent one of the fastest-growing fields in solid-state physics. For one thing, the semiconducting materials being made into compositional superlattices now go well beyond gallium arsenide and aluminum gallium arsenide. One of the most intriguing new crystals is surely the compositional superlattice made of layers of indium arsenide (InAs) and gallium antimonide (GaSb), which was proposed and first investigated by George A. Sai-Halasz, Chang, Tsu and Esaki at IBM. Peculiarities in the electronic properties of this superlattice arise from the unusual positions of their bands. It happens that the lower edge of the conduction band in indium arsenide is below the upper edge of the valence band in gallium antimonide. Thus there are energy levels at which electrons and holes coexist, the former in the indium arsenide, the latter in the gallium antimonide. The coexisting carriers mix their characteristics and are becoming known as hoctrons.

Solid-state superlattices are far more than elaborate toys for playful solid-state physicists. In my opinion compositional superlattices, and particularly doping superlattices, with their unique tunability, represent an important new class of semiconductors, one that will have quite an impact not only on solid-state physics but also on future electronic technology.

The Social Archaeology of Megalithic Monuments

The change from simple tombs to elaborate henges in the Neolithic period of western Europe appears to have coincided with the rise of centralized political control

by Colin Renfrew

Over the past two decades many prehistorians have moved beyond the concept of re-creating the past in terms of culture history toward a concept of "process" that seeks to explain past events rather than being content to narrate them. As Kent V. Flannery of the University of Michigan has put it, some New World archaeologists are not "ultimately concerned with 'the Indian behind the artifact' but rather with the system behind both the Indian and the artifact: what other components does the system have, what energy source keeps it going,... and so on?"

The distinction between the two approaches to prehistory is made clear in the Old World by the case of the European megaliths (from the Greek *megas,* "great," and *lithos,* "stone"). These impressive monuments have intrigued students of the past for more than a century. They are found in all the countries of Atlantic Europe, from the Mediterranean seaboard to Sweden, and the human bones many of them contain show that they served as tombs. Some are very simple, although they could not have been easy to build. For example, the monuments called dolmens may consist of three or four large boulders supporting a massive capstone. There are some 50,000 dolmens in Europe, and the labor required to group the boulders, to build up earth ramps and then to move the capstones into position must have been prodigious. Other monuments are far more elaborate than dolmens and must have required man-hours of labor totaling in the hundreds of thousands. Examples include the great passage graves of Newgrange in Ireland and of Maes Howe in the Orkney Islands.

Even the early students of the megalithic monuments of Europe realized that they were pre-Roman and indeed prehistoric. The scale and sophistication of such structures as the passage graves, however, made the same scholars reluctant to believe that the megalithic monuments were the unaided work of the barbarian peoples of European prehistory. The absence of metal objects in them indicated that their makers were the simple farmers of the New Stone Age, but before the development of absolute dating techniques their actual age could only be estimated.

Did the structures represent the influence of more advanced ideas, percolating from areas that supported high cultures, such as Crete and Greece, westward to Spain and then northward along the Atlantic coast? Given such an assumption, the "diffusionist" view that barbarian Europe merely mirrored the civilizations of the Near East was quick to rise. Scholars of the culture-history school even mapped the successive advances of the supposed Mediterranean influences.

The development of carbon-14 dating caused the complete collapse of the diffusionist view. It was soon learned that the megalithic monuments in several areas of Europe were nearly 2,000 years older than any of their supposed Mediterranean predecessors. Some of the earliest of them, in Brittany, could be dated back to 4500 B.C. There could no longer be any doubt that the monuments, the earliest stone structures still standing anywhere in the world, were of local European origin.

The dual problem that then faced the archaeologist was not only to find some explanation for the origins of these monuments but also to give some account of their function that could make the monuments intelligible to us. The various historical reconstructions that were once offered are now seen to have failed. Instead one has to think in terms of process, considering the system behind both the monuments and the societies that built them. Today archaeologists still have only a few clues to why the megalithic monuments rose precisely where they did, and not elsewhere in Europe, but at least one can begin to think about them in social terms.

Although the structures may have had other functions, most of them served as tombs. In some areas, notably in Britain, there are prehistoric stone monuments of other kinds, notably the great stone circles of Stonehenge, Avebury and the Ring of Brogar in the Orkneys. In Brittany great "alignments" of standing stones arrayed in parallel rows three-quarters of a mile long are found at Carnac. Carbon-14 dating and the variety in the form of the monuments in different areas suggest that they may have had five or six quite independent places of origin. One of those places was surely Brittany. Another was in Portugal and perhaps in Spain; others were in Denmark, probably in Ireland and perhaps in southern England. Glyn Daniel of the University of Cambridge has suggested that some of the monuments took their form through the imitation in stone of the wood houses of the local inhabitants and their ancestors. This is in some instances a plausible suggestion, but it does not state precisely why the monuments came into existence.

Faced with such a problem, the archaeologist would do well to ask what the role of these diverse monuments was in the societies of the time. This is in part a question about the actual function of the monuments: How were they used, how did they facilitate the workings of the society? In part it is a further question about the nature of those societies. What kind of societies were they in order for such monuments to have a meaningful place in them? And of course both questions go beyond the issue of "usefulness" in any narrow materialist or "functionalist" sense; it is evident at the outset that these great feats of construction must in many instances have had an enormous symbolic value. They were surely a source of pride to their makers and perhaps of envy to their neighbors. In a social approach to



DITCH SURROUNDING THE RING OF BROGAR, one of two henge monuments on the central Orkney island known as Mainland, was trenched in the course of the author's studies there. It proved to be the result of quarrying in the sandstone bedrock. The labor investment required to complete the late Neolithic ditch, more than 330 meters in circumference, is estimated to have been 80,000 man-hours.

these questions the human qualities of pride and emulation, of solidarity and competition must have a role.

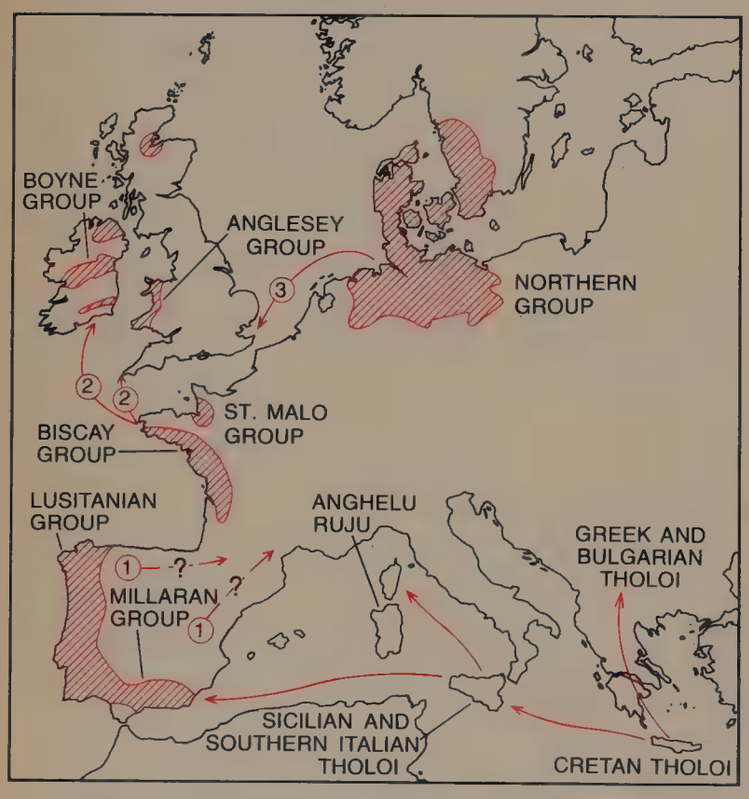
Whatever their general interest, the two questions must in practice be asked first about a specific region and place. The monuments differ from area to area, and so must the specific interpretive approach, whatever regularities may underlie them. One of the first areas to be examined in this way was prehistoric Wessex: those several counties of southern England where Stonehenge, Avebury and many lesser monuments are found. The period is between the first development of farming communities there in about 4000 B.C. and the development of a technology based on the working of bronze in about 2000 B.C., when the building of megalithic monuments had effectively gone out of fashion. Until recently little was known about the domestic settlements of this area. Even today most of our knowledge comes from the considerable profusion of monuments, many of them on the rolling chalklands that were among the first areas to be settled.

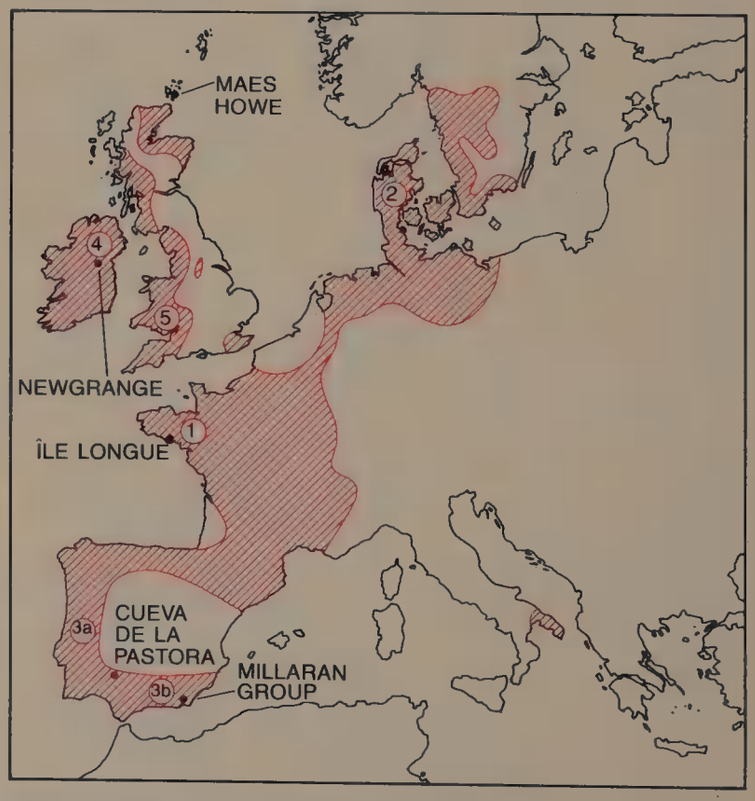
The earliest monuments in Wessex are not in fact megalithic: they have no large stones. They are nonetheless burial monuments. These great oblong mounds, up to 70 meters long, are known as long barrows. Made of chalk, they are in some ways the local equivalent of the stone-built tombs commonly found to the west and, more sparingly, in the Wessex area itself. Because early investigators did not find tomb chambers in them, the structures were at first called unchambered long barrows. Later excavations, however, revealed that most of them did originally contain wood structures that had long since collapsed. Like megalithic chambered tombs, the barrows undoubtedly served as places of burial, but the number of individuals represented by the remains in them is in most instances small. Moreover, the remains are incomplete, suggesting that the dead had been allowed to decompose elsewhere in preparation for secondary burial in the barrow (a custom that is well documented for cultures in many parts of the world).

Grave goods are sparse in the long barrows, as they are in nearly all the European megalithic tombs, and what is there is very simple. The finds are generally restricted to the normal range of Neolithic artifacts: pottery, polished stone axes and chipped-flint tools. Except in the tombs of Spain and Portugal hardly any attempt was made to immure richly decorated objects such as stone carvings or pieces of incised or otherwise decorated stone. It is thus fair to say, both for Wessex and a larger area, that the tombs did not serve for displays of personal wealth, nor do the objects included in them indicate that the dead were of high social status. This contrasts markedly with the finds from the succeeding early Bronze Age in Wessex. In the centuries immediately after 2000 B.C. individual burials were accompanied by rich grave goods, including objects of bronze, amber and gold.

A second class of stoneless Wessex monument is known as a causewayed camp. Most of the camps are roughly circular enclosures about 200 meters in diameter surrounded by concentric rings of ditches. The ditches are not continuous; they are broken by undug areas (the causeways), and it is now clear that the banks thrown up with the chalk excavated from the ditches, rather than the ditches themselves, are what were originally significant. Recent excavations at Hambledon Hill in Dorsetshire by Roger Mercer of the University of Edinburgh have shown the enclosure to have been defended by outlying timber palisades on the approaches to it. Moreover, Mercer found evidence, in the form of human skulls and skeletal fragments in the ditches, that amplified earlier interpretations of these sites as the meeting places of scattered communities from the surrounding areas. Mercer suggests that the enclosure served as a "corpse exposure center, . . . a vast, reeking open cemetery, its silence broken only by the din of crows and ravens."

Both the long barrows and the causewayed camps belong to the earlier part of the British Neolithic, from about 4000 to 3500 B.C. Later, in about 3000 B.C., a different kind of monument made

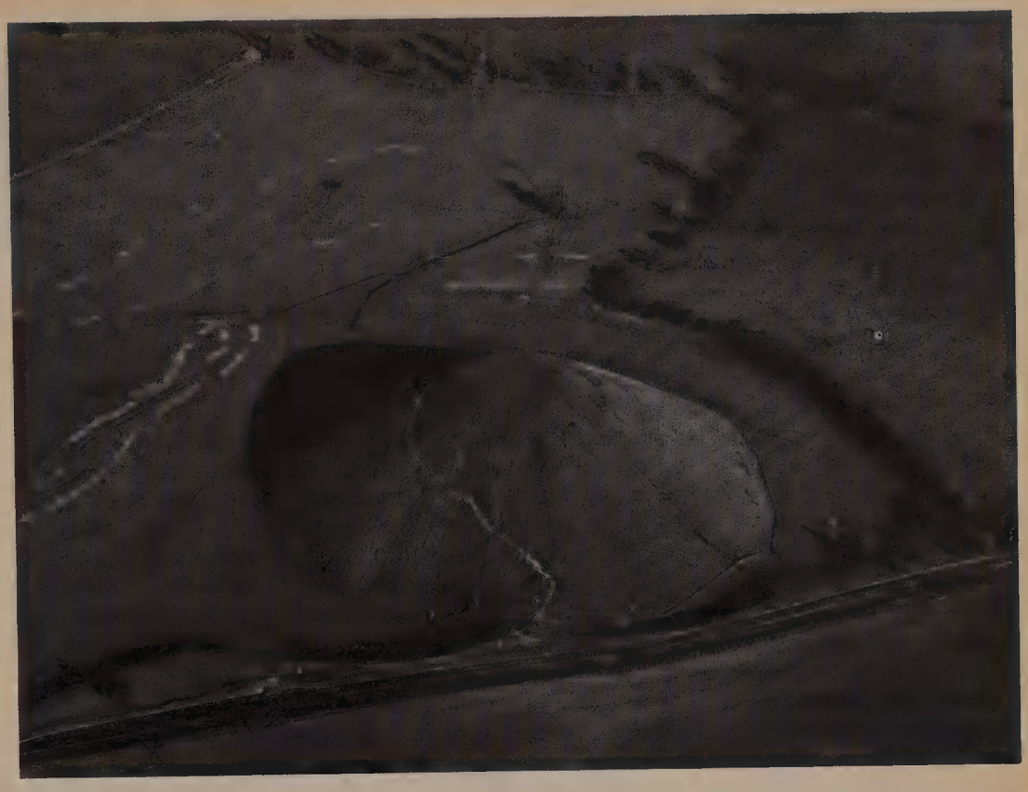




SUPPOSED DIFFUSION of the megalithic tradition in Atlantic Europe, as mapped in the days before carbon-14 dating, was based on the assumption that the beehive tombs of Crete and Greece, with their corbeled, earth-covered roofs, were older than similar tombs in Italy and Sicily and that the latter in turn had inspired such premegalithic oddities as the rock-cut tombs of Malta and Sardinia. Thereafter diffusion of the megalithic tradition through the Iberian peninsula to western France, northern Germany and parts of Scandinavia and finally to England, Wales, Scotland and Ireland would have followed (*map at left*). With the development of absolute dating, which showed that certain French megalithic monuments had been built in about 4500 B.C. and that even the monuments in the remote Orkney Islands were centuries older than the pyramids of Egypt, the diffusionist view gave way to one emphasizing local origins (*map at right*).

its appearance: the henge, a term derived from the famous Wessex monument Stonehenge. Most henges are not, however, a circle of great stones. They are a circular enclosure surrounded by a single ditch with a well-constructed bank that is usually on the outside of the ditch. There are generally either one or two entrance causeways, and the entire arrangement is much more regular than the earlier causewayed enclosures. Excavations at a number of henge sites (notably those at Durrington Walls undertaken by Geoffrey Wainwright, Inspector of Ancient Monuments for the U.K. Department of the Environment) have shown that some of them included massive circular timber buildings. Henges are sometimes very large, up to 500 meters in diameter.

Finally in this brief review of the Wessex monuments mention must be made of Stonehenge itself, with its circle of stones linked together by lintels, and of the colossal artificial earth mound, Silbury Hill, near a stone circle far larger than Stonehenge: Avebury. Until the time of the Industrial Revolution, Silbury Hill ranked as the largest manmade structure in Europe.



SILBURY HILL, a large artificial earth mound near Avebury in Wiltshire, appears at the center of this aerial photograph. It was the largest manmade structure in Europe until the time of the Industrial Revolution, requiring an estimated investment of 18 million man-hours to build.

In trying to make sense of such a profusion of material two analytical approaches are potentially useful. The first approach takes into account spatial distribution. The second considers absolute scale. Both approaches may give indications of hierarchy, and this in turn may reflect an underlying social ranking within the society itself. For example, when the Wessex monuments are mapped, it becomes clear that the 132 long barrows outnumber the seven causewayed camps by nearly a factor of 20. The map also shows a dispersed distribution of long barrows. They fall into perhaps five groupings, and within each of the groupings is a convenient causewayed camp. (The two camps left over do not fit as well among the long barrows.) Thus it is already possible to speak of a modest spatial hierarchy if one regards each causewayed camp as a local center for the small region defined by each group of long barrows. When one goes on to look at the distribution of the major henges (those more than 200 meters in diameter), a spatial pattern emerges that is similar to the pattern of the causewayed camps. Each causewayed camp, again with two exceptions, is superseded in this later period by a major henge.

The analysis is taken further when we turn from spatial distribution to scale. For example, the long barrows are of rather modest size. Calculation of the amount of labor required for their construction (following estimates based on simple earth-moving methods and the use of very basic tools, such as wood shovels and antler picks) suggests that each would have taken 5,000 to 10,000 man-hours to build. This is an investment of labor with an order of magnitude of from 10^3 to 10^4 man-hours: the amount of work 20 men could do in 50 days. The labor required for the causewayed camps, next in hierarchical rank, is an order of magnitude greater. They would have required between 40,000 and 100,000 man-hours, that is, up to about 10^5 man-hours.

The great henges of the late Neolithic represent a labor investment again an order of magnitude larger, approaching 10^6. For example, Wainwright has estimated a total of 900,000 man-hours for the construction of Durrington Walls. And if the great henges rank as superstars among prehistoric monuments, Wessex provides two megastars: Silbury Hill and Stonehenge. Richard J. C. Atkinson of University College, Cardiff, has estimated the labor requirement for Silbury Hill and Stonehenge (including the labor necessary for the transport of the stones to Stonehenge) at 18 million and 30 million man-hours respectively. This is well in excess of 10^7 man-hours.

There is thus a man-hour hierarchy to set beside the evident spatial hierarchy. Some years ago I suggested that we should postulate social structures correlating with these space-cum-energetics hierarchies and that, for the late Neolithic, it might be appropriate to talk of "chiefdom" societies centered on the five very large Wessex henges. The approximate territories can be inferred by dividing up the landscape with each of the major henges as the dominant focus of its region. I pointed out that for so great an investment of labor it is necessary to think of some central organizing authority and that it is possible to list a number of relevant features of chiefdom societies as ethnologists recognize them. Fundamental among these features, of course, is the existence of a ranked society (with a chief at its head) sustained by a system based on a redistribution, organized by the chief, of produce, including some of the necessities required for subsistence. Chiefdoms characteristically have a higher population density and more clearly defined territorial boundaries than more egalitarian societies. They have centers that coordinate social and religious activities as well as economic ones, and they show an increase in specialization, as well as the ability to organize and deploy labor for major works.

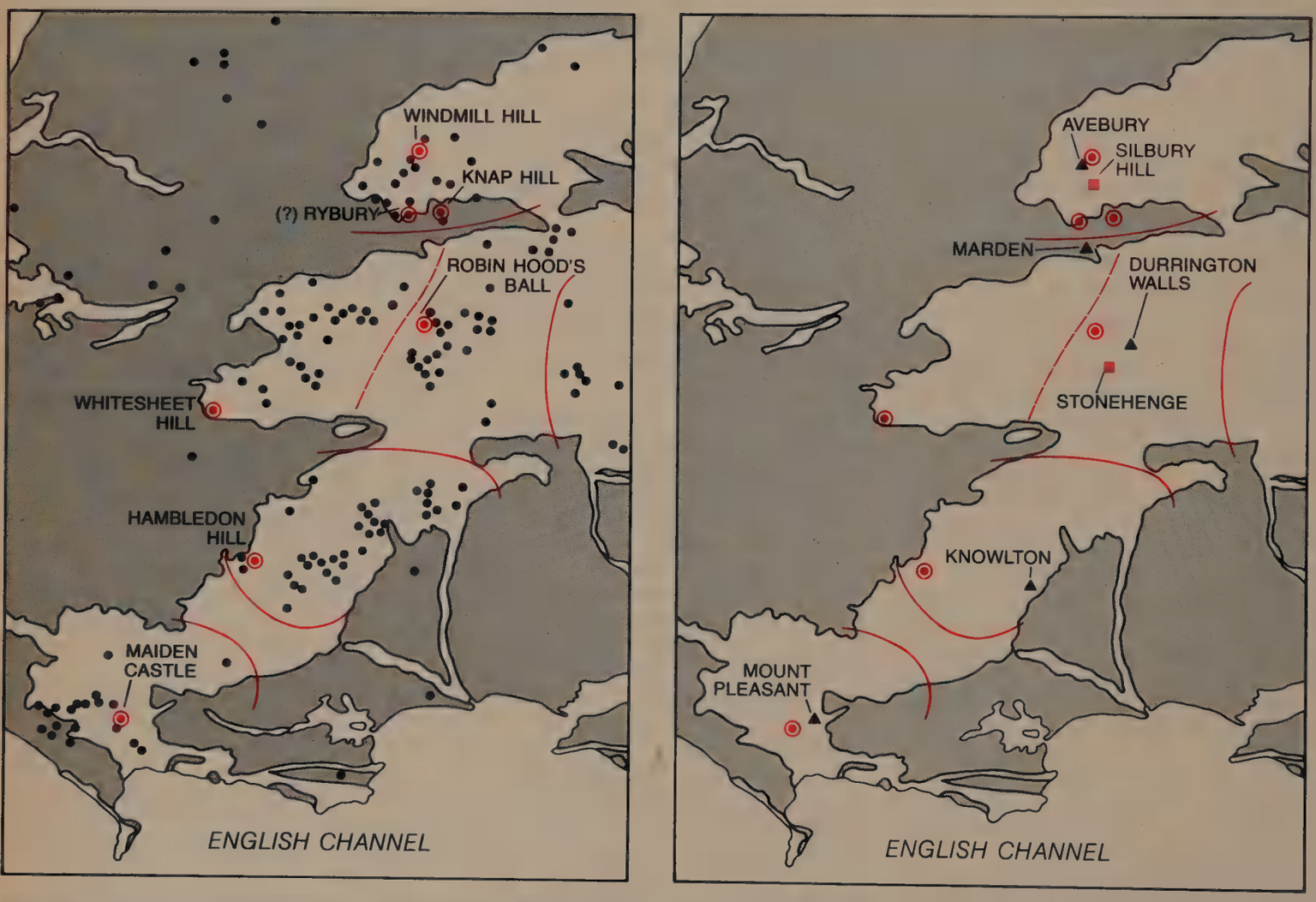
Some of these features might already be inferred for polities with sites like Durrington Walls or Avebury at their center. As the next step upward some further centralization would have arisen: an amalgamation or association of formerly independent chiefdom territories into a federation, perhaps with Stonehenge as its ritual focus and Silbury Hill as its largest monument. On the basis of his excavations at Silbury Hill, Atkinson has suggested that the monument was built within the space of only two years. Even with workers laboring eight hours a day for 300 days a year a full-time team of 3,700 men would have been needed to construct the great tumulus, and each member of the team would have needed to be sup-

plied with food by the organization of his respective chiefdom.

Such a "social archaeology" approach calls for thinking in terms of a well-defined social context for the larger monuments. It situates them within a dynamic development where the simpler societies of the early period developed into the more centralized social units of the later period, which should have had a greater population density and hence a larger population. It is an approach that also allows one to separate the question of the larger monuments (the great henges and Silbury Hill), which we can regard as the product of chiefdom societies, from smaller local monuments such as long barrows. This is important because the large majority of the European megalithic monuments can be more appropriately compared with long barrows. They too are burial monuments, often requiring about 10,000 man-hours for their construction, and many of them are as dispersed as the long barrows. Major monuments are much rarer: henges and stone circles are hardly found on the European mainland (although, at least as far as scale is concerned, the great alignments at Carnac undoubtedly belong in the Stonehenge category).

Concepts of scale and spatial organization are again useful whether one looks more closely at the long barrows of Wessex or at stone monuments of the same antiquity, such as the megalithic tombs on the island of Arran off the west coast of Scotland. In either case if one draws lines midway between the monuments to divide the landscape into separate units (a process geographers call forming Thiessen polygons), one finds that with few exceptions the landscape is divided into roughly equal productive units. For example, on Arran each such polygon contains at least some of the island's limited arable land. Setting aside the instances where the monuments are found in pairs, the pattern is also a dispersed one. If in addition one takes into account the size of the monuments, there is still no sign of a hierarchical structure: at least within one order of magnitude the tombs are approximately the same size. The pattern is not chiefly but egalitarian.

The obvious suggestion arises that these monuments of 5,000 or 6,000 years ago serve as indicators of small, rather scattered groups of farmers living in societies that were not centrally organized. No doubt these groups were linked by marriage and hence by kinship ties (and perhaps also by exchanges of goods), but they may well have been politically independent. Tribal ties between them, initially on a rather modest scale, may be indicated by the next phase: the construction of causewayed enclosures. We may imagine the dead of all the 20 or so long-barrow territories in the Hambledon Hill region, so vividly described by Mercer, being brought



BURIAL MONUMENTS of ancient Wessex (modern Dorsetshire and Wiltshire) in England form clusters that are concentrated largely in the unwooded chalk uplands (*white areas*) that were preferred by the Neolithic farmers who first settled the region. Their burial mounds (*dots*) were long earth barrows and therefore not megalithic monuments at all. When they are grouped into five or six clusters (*colored boundary lines*), however, each of the four lower clusters contains a single causewayed camp (*colored bull's-eyes*), which seems to have been an area where the dead were first exposed before their bones were finally interred in the associated group of long barrows (*map at left*). The grave goods found in the long barrows are few and simple, suggesting that the farmers formed egalitarian social groups. When, about 500 years later, new kinds of circular henge earthworks that may all have included timber structures within their bounds were built in this same part of England, several of the larger ones (*triangles on map at right*) lay within the same bounds of clustered long barrows as the earlier causewayed camps did. The henge monuments, which rarely include megaliths, are substantially more elaborate than the earlier causewayed camps. This suggests that the social organization they served had become more hierarchical. Further evidence to the same effect is the appearance of two "megastar" monuments in the region (*rectangles*): Stonehenge and, to the north, Silbury Hill.

there for exposure and the bones of some of them being later taken back to the local long barrow for interment. Since the long-barrow territories are about 10 square kilometers in area, taking a modest figure of between 10 and 50 hectares of land to support each person in a simple farming economy, we may think in terms of populations of between 20 and 100 people per territory. These small units may be regarded as "segmentary societies," to use the social anthropologist's term, each segment being a group of people, an economically and politically autonomous, self-sustaining perpetual body exercising effective control over its productive resources.

From this perspective we can begin to think more clearly about the far more ambitious megalithic chambered tombs. They were, of course, burial places. Indeed, with the possible exception of the timber-chambered long barrows, they were collective burial places. But a group of perhaps 50 people, even a permanent group occupying the same piece of land for centuries, does not actually need a monumental tomb to dispose of its dead. There would be nothing simpler than to bury them quickly, without any fuss. The monumentality of the chambered tombs, therefore, is not logically implied by their funerary function. It hints at another purpose altogether. In the words of Andrew Fleming of the University of Sheffield, these were "tombs for the living," that is to say, their monumental character is the result of the deliberate activities of the living members of the society and must have served the objectives of that society. Three different explanations can be put forward for the role of the chambered tombs. The three are not contradictory but serve rather to complement one another.

First, with respect to these ambitious monuments, I have suggested a function in terms of what may be called social cohesion. In times of stress, perhaps particularly in circumstances of high population pressure, a community could function more effectively when its social unity was emphasized and felt by all. The building of such an enduring monument was a major symbolic act asserting just this unity. It should be remembered that these people lived in a world where other obvious lasting human creations were few: even the houses were usually made of wood and other perishable materials. Once the monument was built, however, it would be the major manmade landmark of the territory. Sometimes there must even have been competition between neighboring territories to have the largest and most conspicuous center.

I argue, then, that we should view these monuments not merely as tombs but as public centers. Often they might have served as meeting places: perhaps the locus for an entire range of religious rites relating the community as a whole to its ancestors as well as to its more recent dead. The building of a focal monument could have given visible expression to the communal aspirations of the local group, serving to ensure the group's continuing cohesion and hence its survival.

Second is a closely related idea developed by Arthur A. Saxe of Ohio University, who has pointed out the relevance of such funerary activities to a community's maintenance of its right to ancestral lands. As he puts it: "To the degree that corporate group rights to use and/or control crucial but restricted resources are attained and/or legitimized by means of linear descent from the dead (i.e., lineal ties to ancestors), such groups will maintain formal disposal areas for the exclusive disposal of their dead." Title to communal land is thus asserted by the maintenance of the ancestral tomb. Both views may be summarized as saying that the monuments served as the territorial markers of segmentary societies.

Third, Christopher Tilley of the University of Lund has taken a neo-Marxist approach to suggest a further function of the monuments, one that does not contradict the two preceding ideas. Tilley stresses the role of mortuary ritual, in common with other life-cycle rituals, in "legitimizing sectional interests" within the society. Even within the relatively egalitarian framework of a segmentary, lineage-organized society, he suggests, there are class contradictions, notably between the elders of the group and the junior members. He argues that control over ritual, emphasized by such monuments, helped to facilitate the continuing dominance of the elders. I am not sure that such an argument takes us much further than seeing the monuments as symbolizing and facilitating social cohesion, as has already been suggested, but Tilley certainly sets this idea in a different framework.

Each of the three explanations is functionalist. That is to say no one of them effectively explains the origin of megaliths but each shows how the monuments played a useful and meaningful role in the societies that erected them. This point of meaningfulness must be at least one component of any satisfactory explanation for them, and these are certainly explanations in social terms.

Such ideas, of course, are difficult to test directly in the way that one would wish to test a generalization or a hypothesis in the exact sciences. They can nonetheless be examined afresh in the light of new work in the field. It was with this intention that between 1972 and 1974 I undertook field work in the Orkney Islands at the northernmost tip of the British Isles.

The Orkneys are virtually without trees and have been since the last ice age because of the prevailing strong winds. The local rock is an easily fractured laminar sandstone, so that the prehistoric population had readily available an abundant building resource other than timber. All Orkney buildings were therefore built of sandstone, and they have survived particularly well. Until very recently farming practices on the islands were not mechanized, and so the cairns of stone that are today the indicators of buried prehistoric remains have not been demolished.

The prehistoric antiquities of the Orkneys have long been famous, notably two henge monuments with circles of standing stones, the settlement at Skara Brae and the splendid passage grave of Maes Howe. Skara Brae, excavated in the 1930's by V. Gordon Childe, is a magnificently preserved village from the time of the island's early farmers (about 3000 B.C.). The walls of the houses still stand to a height of up to two meters, and some internal furnishings such as beds and cupboards, being also built of stone, have survived. Maes Howe has an impressive stone-built chamber entered by a passageway. The stonework is of such quality, the stones having been carefully fractured, that only with difficulty can one accept that no metal tools were used in its construction. On the old diffusionist theory for the origin of the megaliths, Maes Howe, being the finest monument in the Orkneys, was also considered to be the oldest: the first inspired effort of the supposed new immigrants as they landed with their civilized ideas still unimpaired by decades of living in the remote north.

At the time I began work, no carbon-14 dates at all were available from the Orkneys, and one of my aims was to obtain datable samples of organic material relevant to the chronology of the Orkneys chambered tombs. The main site chosen for excavation was the tomb of Quanterness. From a report written early in the 19th century I knew that Quanterness had been entered at that time and had then been closed. My colleagues and I hoped to find evidence for its date of construction and by conducting further studies to contribute to the understanding of the islands' monuments in general. At about the same time fresh excavations directed by David Clarke of the National Museum of Antiquities of Scotland were beginning at Skara Brae, and Graham Ritchie of the Royal Commissions for Ancient and Historical Monuments of Scotland was starting work at one of the Orkneys henge monuments, the Stones of Stenness.

The roof of Quanterness had collapsed, and we had to enter the chamber

from above. As the 19th-century sketch plan indicated, there were six side chambers leading off the rectangular main chamber, with an entrance passageway from the east, now blocked. The early account had not, however, prepared us for the handsome, regular stonework, which makes this one of the most impressive buildings of its period in existence anywhere. Nor had we expected the tangled profusion of human and animal bones we encountered once the debris had been removed from the upper levels. The floor of the main chamber was covered with disarticulated human remains, strewn around in great confusion, together with fragments of pottery and other objects. It was necessary to use a water-sieving technique on all the material from the tomb in order to extract small fragments of artifacts and of bone, including the bones of fishes and birds.

The clearance of these main-chamber finds enabled us to enter the side chambers of the tomb. Four of the six were intact. Once the connecting passages were cleared it was evident that these side chambers were not blocked; it was possible to bend down, crawl in and stand upright in a perfectly complete stone chamber, beautifully roofed, that had existed for more than 4,000 years. Carbon-14 dating later revealed that the tomb had been in service from about 3200 B.C., and so the structure was fully 5,000 years old: older by some centuries than the oldest pyramids of Egypt.

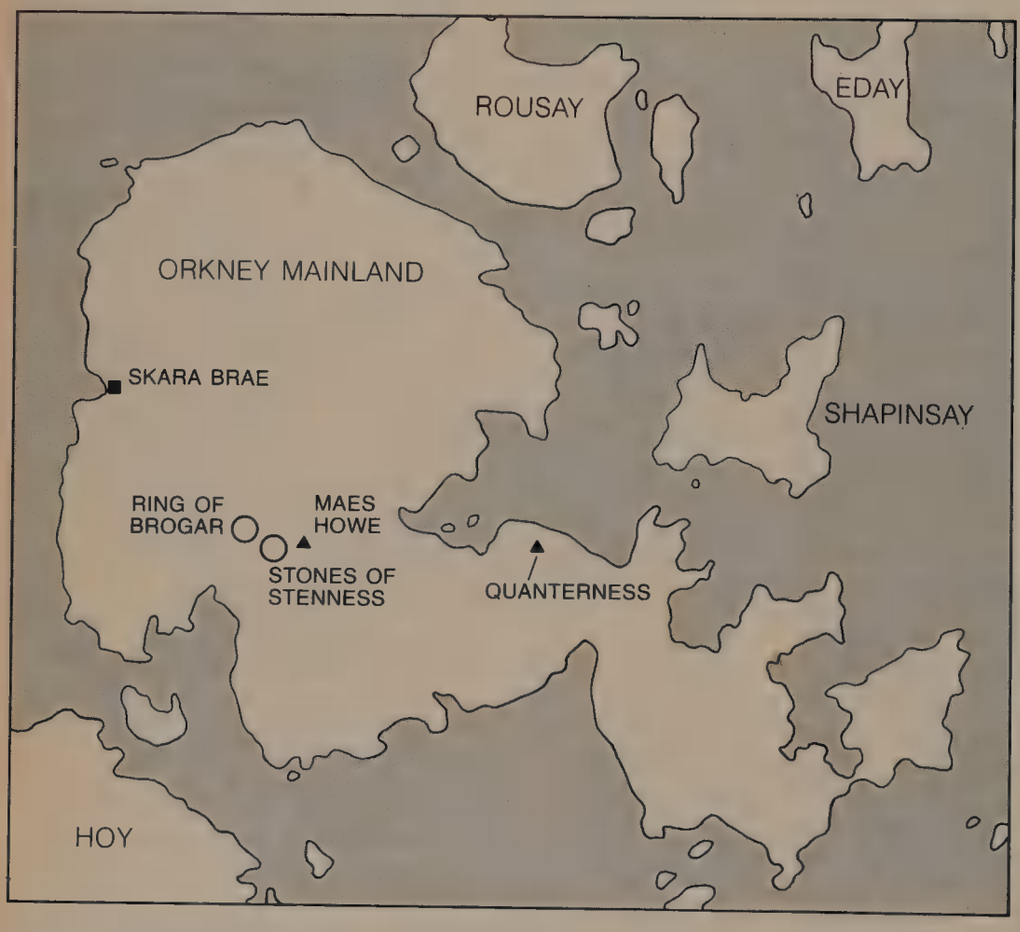
There were few grave goods in the tomb. We found the fragmentary remains of some 34 pots, only three polished stone knives and an occasional piece of chipped flint. The animal bones, clearly the remains of joints of meat that had been brought into the tomb either as offerings to the dead or as food for consumption during funeral rituals, represented parts of a minimum of seven sheep, 18 lambs and five oxen. Also present were the bones of red deer and other wild mammals, of at least 35 birds of various species and of seven species of fishes.

Analysis of the very fragmentary human bones by Judson Chesterman of the University of Sheffield produced surprising results. From the sample of 12,600 pieces we recovered he was able to estimate that the partially excavated main chamber of the tomb and the one side chamber we fully examined held parts of at least 157 individuals. As in the Wessex long barrows, the dead had clearly been exposed before their bones were collected for final burial in the tomb. If this sample figure is increased by a suitable factor to allow for the 20 percent of the main chamber and the other five side chambers that remain unexcavated, Quanterness probably holds the remains of no fewer than 400 individuals.

The ratio of males to females (54 : 46) is sufficiently close to equality to suggest no exclusion from the tomb on the grounds of sex. With the exception of infants all age groups are represented. The age distribution shows that the mean life expectancy was between 20 and 25 years. Although such a life span is short, it is not out of keeping with finds from other Neolithic sites, and it suggests that what we uncovered may have been the dead from an entire living community. From the carbon-14 estimate of how long the tomb was in service, about 550 years, we can place the size of that community at between 13 and 20 individuals.

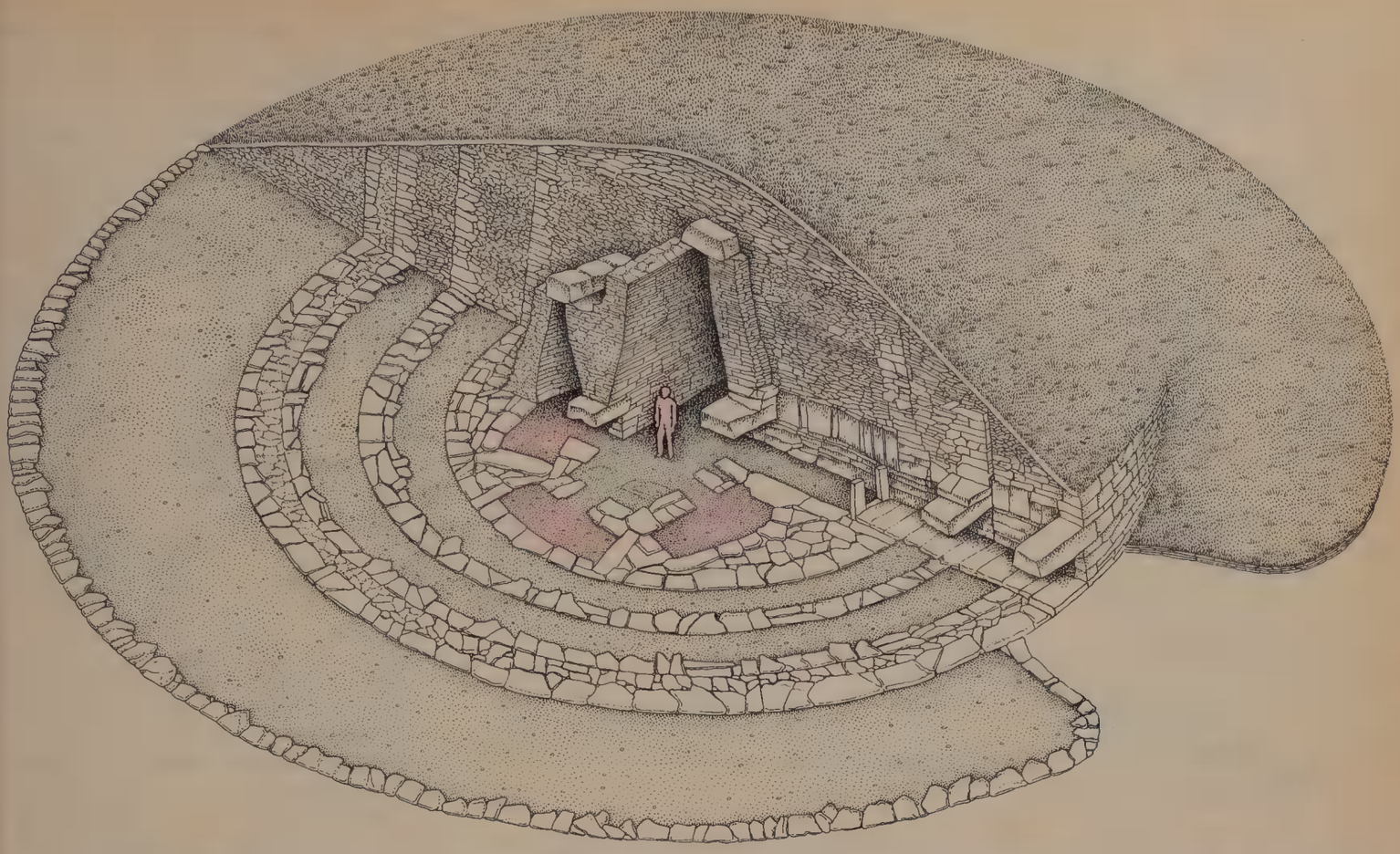
While excavations were continuing at Quanterness we were able to examine some other Orkney sites. Calculation showed that Quanterness represented a labor input of about 7,000 man-hours for the tomb and its enclosing cairn, although this figure takes no account of the skill of the builders. It is clear that most of the other Orkney chambered tombs, which have a notably dispersed distribution, represent labor inputs of the same general order of magnitude.

Our work at the other henge, the Ring of Brogar, showed that the ditch surrounding the henge was a product not of earth-moving but of splendid rock-cutting, representing a labor investment of about 80,000 man-hours (not taking into account the erection of the standing stones of the circle). Unfortunately we were not able to date the monument satisfactorily. The best estimate of its age comes from the carbon-14 date, about 3000 B.C., obtained by Ritchie for its smaller sister monument, the Stones of Stenness. We were able to obtain dates from the ditch surrounding the great tomb, Maes Howe, which suggest that it was built in about 2800 B.C. Because the construction stone had to be brought from a considerable distance, Maes Howe represents a labor input more like that of the Ring of Brogar than that of Quanterness.



FOUR MEGALITHIC MONUMENTS, two henges and two chambered tombs, are among the eight found on the largest of the Orkney Islands. The author and his colleagues reopened the chambered tomb Quanterness, built in about 3200 B.C., for the first time in nearly two centuries and partially excavated it. They also trenched the perimeter of the Ring of Brogar, a henge monument probably put up in about 3000 B.C., perhaps two centuries before the more elaborate chambered tomb of Maes Howe was built. The simplicity of Quanterness, compared with the other structures, suggests that the farmers who built it formed one small community.

These various results allow us to draw conclusions of a social nature about early Orkney that can be compared with the more general social-archeological picture outlined for Wessex. First, at the time of its early use the tomb at Quanterness was not stratified spatially above or below monuments built on a different scale. There simply was no such spatial hierarchy. Second, Quanterness was an "equal access" tomb, containing a balanced representation of both sexes and of all ages except for infants. Third, the



QUANTERNESS TOMB is shown in this cutaway drawing as its northern half probably appeared soon after construction. The human figure at the center (*color*) gives the scale; the outer stone perimeter is 45 meters in circumference and the tomb's height at the center was originally some 3.5 meters. Three of the six side chambers can be seen in the cutaway (*light color*); the author and his colleagues excavated one of the side chambers completely and partially cleared the main chamber. The estimated total number of burials is 400.

grave goods were very simple, indicating no prominent ranking among the occupants of the tomb. This finding, along with the equal-access one, supports the view that the people belonged to a relatively egalitarian society. Fourth, burial practices were elaborate: inhumation of the bones within the cairn was the last stage in the treatment of the deceased, having been preceded by the decomposition of the corpse. Finally, the chamber was in use for at least five centuries.

Let us now consider still other facts that we learned bearing on the nature and sequence of events at the site. Item: The group that used the cairn probably numbered no more than 20, including men, women and children. Item: The labor required for the construction of the monument, less than 10,000 man-hours, could without difficulty have been invested by such a group over the span of a few years, or perhaps over an even shorter time if the builders had the assistance of neighboring groups. Item: Quanterness is just one of a number of similar Orkney burial cairns of comparable scale. Item: The distribution of these cairns is fairly dispersed, suggesting that the group making use of each cairn occupied and farmed the surrounding territory. And lastly, some centuries after the construction of Quanterness a few monuments built on a far larger scale, including the Ring of Brogar and Maes Howe, were raised in a single central area of the same main island of the Orkneys. The new construction can be taken to imply the emergence of some form of more centralized social organization.

The finds from Quanterness and nearby sites thus lead to, from excavational data, conclusions that harmonize well with those already arrived at Wessex. Whereas the time elapsed between egalitarian and centralized societies is not long enough in the Orkneys to demonstrate this sequence of social development conclusively, it does seem clear that the raising of the major monuments there, indicative of a greater degree of centrality, was relatively late in emerging and that most of the cairns belong to an earlier, more egalitarian phase.

The abundant skeletal remains from Quanterness help to confirm the view that these cairns were not built just to house one or two wandering missionaries. Quite simply they were erected by the local inhabitants: a relatively dispersed farming population. The considerable sophistication of their architecture and their skill in construction both point to the great pride taken by the parent community in its monument.

It is not argued here that the monuments of the Orkneys had an origin independent of the other areas where collective burial in monumental tombs was practiced. I suggested above that there may have been five or six primary megalithic centers in western Europe, including one in Ireland and perhaps still another in southern England. The spread of farming techniques to Scotland is likely to have been from one or another of these areas, and the colonists probably brought the custom of tomb-building with them. The first inhabitants of the Orkneys in turn must have crossed to the islands from northern Scotland and were no doubt familiar with the rather simple megalithic tombs constructed in the north at the time. What was special about the Orkneys was the impressive development of a local architecture, based on the excellent local sandstone, to give such masterpieces of prehistoric architecture as Quanterness and Maes Howe. And the particular interest of the Orkneys to the archaeologist is the great abundance of prehistoric remains, which makes the attempt to think coherently about these considerable achievements in social terms less difficult.

The evidence from the Orkneys adds significantly to the picture one gets from Wessex. It is possible to suggest what the function of these monuments may have been, at least in some of the societies that built them, but if one is now to go on to explain megalithic origins, one must also indicate why they

INVEST YOURSELF



A windmill to pump water for "salt farming" in India. More efficient woodburning stoves for the Sahel. Photovoltaic irrigation pumps for the Somali refugee camps

All these are solutions to technical problems in developing countries. Devising such solutions is no simple task. To apply the most advanced results of modern science to the problems of developing areas in a form that can be adopted by the people requires the skills of the best scientists, engineers, farmers, businessmen—people whose jobs may involve creating solid state systems or farming 1000 acres, but who can also design a solar still appropriate to Mauritania or an acacia-fueled methane digester for Nicaragua.

Such are the professionals who volunteer their spare time to Volunteers in Technical Assistance (VITA), a 20 year old private, non-profit organization dedicated to helping solve development problems for people world-wide.

Four thousand VITA Volunteers from 82 countries donate their expertise and time to respond to the over 2500 inquiries received annually. Volunteers also review technical documents, assist in writing VITA's publications and bulletins, serve on technical panels, and undertake short-term consultancies.

Past volunteer responses have resulted in new designs for solar hot water heaters and grain dryers, low-cost housing, the windmill shown above and many others. Join us in the challenge of developing even more innovative technologies for the future.

VITA Putting Resources to Work for People

1815 North Lynn Street, Arlington, Virginia 22209-2079, USA

came about when they did and where they did.

The "when" is not so difficult. These were the monuments constructed by the first farmers in the various regions of Atlantic Europe soon after they had settled down and established a stable way of life in the lands along the Atlantic coast. No monumental tombs of this scale were built by the earlier and much sparser population of Mesolithic hunters, gatherers and fishermen who exploited some of these same coasts before the development of farming.

The "where" is less easy. If the earliest farmers everywhere in Europe had developed similar monuments, there would be less of a problem. Why did the monuments go up in these particular coastal lands and not in either central or eastern Europe? To answer the question one has to isolate some factor special to the Atlantic areas and to show why that factor is relevant.

One such factor is the existence of the sparse but settled Mesolithic population in areas along the Atlantic coast, notably in Portugal and Brittany. We know that these people were already burying their dead systematically in the middens of discarded shells they left behind them. No one has yet explained, however, exactly why the existence of this earlier population should make it more likely that the early farmers would build such monuments.

I propose a different factor and shall do so in demographic terms. It has been suggested that the spread of the early farming economy caused a substantial rate of population increase in every area as it emanated from Greece and the western Mediterranean, until a ceiling was reached at what was for the time a rather high population density. In each instance the approach to this ceiling would necessitate a decline in the rate of increase. Such a decline could not be brought about without some difficulty. In central Europe there was always the possibility of shedding some of the surplus population by means of emigration toward the west and the north, so that the stress of approaching the population ceiling would have been alleviated. In the lands bordering the Atlantic, however, this safety valve was not available. There was simply nowhere to emigrate to, and the demographic consequences had to be faced as best they might.

It is precisely in these circumstances, I would argue, when the population density had become inconveniently high and a need to curb the rate of population increase was evident, that devices favoring good social cohesion within small farming groups and ways of asserting one's right to farming land would be most useful. Communities that had put up monumental territorial markers would have had an adaptive advantage over less assertive and less cohesive neighboring groups. The seeds of these ideas may already have been present in the burial practices of the earlier, midden-depositing population. In any event the unprecedentedly high rate of population increase, impossible to mitigate by an outflow of emigrants, would make useful any devices favoring both stability and social order.

My proposed explanation is only a sketch. It needs further exploration by means of further excavation. For example, an examination of settlements where the earliest farmers may have come into contact with the last of the midden people would be useful. More work and more thought will be necessary before one can maintain either that



JUMBLED BONES, both human and animal, were found on the floor of the main chamber of the tomb of Quanterness. Grave goods were scarce: a few pieces of chipped flint, three polished stone knives and fragments of some 34 pots. Construction of the tomb is estimated to have required some 7,000 man-hours, or about the same labor investment as that for a long barrow.

the problem has been correctly framed or that the mechanism for its satisfactory solution has been found.

If the picture I have sketched can be accepted as plausible, can it be any more than that? Can it be tested with anything like the rigor that one would wish of a scientific theory? The spatial patterns can certainly be tested in principle with the modern geographer's techniques of locational analysis. The real terrain, however, does not usually approximate the level, "isotropic" plain the geographer often assumes, and the ravages of time have caused the disappearance of many monuments. Moreover, it is often difficult to show that the different sites were in use at the same time, as a strict quantitative analysis would require.

Further, it is clear that there can be more than one motive for monument-building, and even in prehistoric times there are notable concentrations of tombs, not dispersed but clustered in cemeteries. Some of the arguments used here regarding dispersed distributions would not apply to the clustered grave sites. In spite of these limitations the archaeologist need not lose heart. Social archaeology, after all, inevitably labors under the same difficulties as the other social sciences. It is dealing with the activities of human beings, which under the best of circumstances are not easy to quantify and are very difficult to predict. That we can deal at all with social issues in a remote past, datable only with the aid of carbon-14, when half the evidence is forever gone, is remarkable enough.

The French historian Fernand Braudel has differentiated three levels at which the processes and events, the "conjunctures" of history, work themselves out. The personal dramas, the "headlines" of the past, are the short-term crises. Underlying these are the economic and social movements that operate over the medium term. Underlying these again is the secular trend, the *longue durée,* where much larger time factors operate. Among them are the almost constant constraints of the landscape and the basic realities of peasant life. In a curious way the archaeologist, who usually can have little to say about the single individual or the short-term event, can sometimes get much closer to the long-term processes than can the modern historian, overwhelmed by the noisy profusion of his evidence for the recent past. That is why archaeologists working on topics of cross-cultural significance, such as the origins of agriculture, the processes of state formation or (as in the present case) the appropriate way of explaining, in a general social perspective, the building of the world's first monuments, feel that in spite of their practical limitations, they have ideas worth discussing.



ANTIQUITY

A Periodical Review of Archaeology
edited by Glyn Daniel

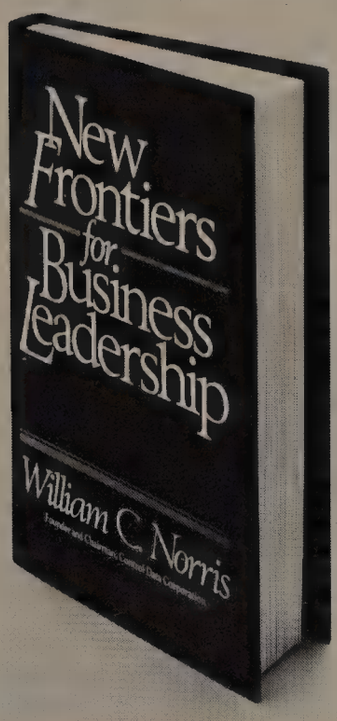
Founded in 1927 by O. G. S. Crawford, ANTIQUITY has appeared regularly ever since and won acclaim the world over as the most authoritative journal in its field. While written by specialists, the articles, notes and reviews are popular in character and indispensable to all interested in the development of man and his past.

Professor Glyn Daniel, Faculty of Archaeology and Anthropology in the University of Cambridge, and Fellow of St John's College, has been Editor since 1956.

The annual subscription, postage included, is $35. Subscription forms and bankers' orders are available on request from

ANTIQUITY PUBLICATIONS LIMITED
Heffers Printers Ltd, King's Hedges Road,
Cambridge, England CB4 2PQ

"Bill Norris ranks as one of the most diversely intelligent and socially committed business executives I have ever known—and this book tells exactly why."
John Kenneth Galbraith



A book that could change the way business does business.

Here is a blueprint for turning the major unmet needs of society into profitable business opportunities. Now available at selected book stores or from Dorn Books, 7831 E. Bush Lake Rd., Minneapolis, MN 55435-3875. (612/835-6855)

Send $15 plus $1 for postage and handling. Visa or Master Charge accepted. Minn. residents add 6% sales tax.

A Seed-eating Beetle's Adaptations to a Poisonous Seed

The beetle larvae feed on a seed containing an amino acid that is severely toxic to other insects. The larvae circumvent the toxic effect of the amino acid and also utilize its nitrogen

by Gerald A. Rosenthal

Most insects depend on plants for their food, and so insects are among the most aggressive and destructive adversaries of plants. Terrestrial plants characteristically lack mobility and cannot elude these destructive pests by flight. Yet plants are far from passive participants in the struggle that has gone on between insects and modern plants for more than 50 million years. The primary weapons of plants against the feeding ravages of insects are the plants' own chemical constituents. The thousands of distinct compounds higher plants synthesize in their metabolism are classified either as primary metabolites or as secondary metabolites. Primary metabolites are substances such as nucleic acids, adenosine triphosphate (ATP) and glucose that are common to all life. Secondary metabolites vary among plants, but plants would be unable to wage chemical warfare without them.

Many biologists believe the feeding activity of insects and other herbivores and the ravages of a host of pathogenic organisms have provided the selection pressure for the elaboration and maintenance of these sophisticated chemical barriers to predation and disease. In such a situation natural selection would have capitalized on the inherent variability in secondary plant chemistry to favor the metabolites that can enhance the organism's Darwinian fitness by affording an effective level of protection. Indeed, in considering the relation between certain insects and plants, Paul P. Feeny of Cornell University has said: "We are thus witnessing an evolutionary arms race in which the plants, for survival, must deploy a fraction of their metabolic budgets on defense (physical as well as chemical) and the insects must devote a portion of their assimilated energy and nutrients to various devices for host location and attack."

The point is aptly illustrated by the plant *Dioclea megacarpa,* a vinelike legume. It has only one insect predator: *Caryedes brasiliensis,* a small beetle of the family Bruchidae, seed-eating beetles of worldwide distribution. One reason the plant is so remarkably successful in repelling other insects is that it stores L-canavanine, an insecticidal amino acid that among other things disrupts the production of normal insect proteins. What is even more remarkable is the adaptation of *C. brasiliensis* to the chemical defense of the plant.

Our work at the University of Kentucky has yielded some insight into the interactions between insects and the secondary metabolites of higher plants. Canavanine is one of about 250 amino acids that are synthesized by higher plants but are not utilized as building blocks of proteins. This nonprotein amino acid is synthesized by members of the Lotoideae (Fabaceae), a group of leguminous plants, and is found in such agronomically important crops as alfalfa and numerous trefoils, including clover. It is also synthesized by annual and perennial ornamental and arborescent species such as *Wisteria* and *Robinia.* There is considerable evidence that canavanine functions as an important nitrogen-storing metabolite, particularly in the seed, where it supports the growth of the newly developing plant.

The molecule L-canavanine is a structural analogue of L-arginine, one of the 20 protein-forming amino acids. In canavanine the terminal $-CH_2$ group of arginine is replaced by oxygen. This difference is of little consequence in the metabolism of canavanine, which therefore serves in virtually every enzyme-controlled reaction for which arginine is the preferred substrate. For example, canavanine is activated by arginyl transfer RNA synthetase, an enzyme that activates arginine and then links it to its transfer RNA. Once canavanine has been linked to the transfer RNA, which normally transports arginine to the protein-assembly sites on the ribosomes of the cell, canavanine is inevitably inserted into the growing protein chain in place of arginine.

Canavanine is chemically much less basic than arginine, and under physiological conditions it would be less positively charged than arginine. This difference in charge can affect the interactions that cause the protein chain to fold up into the uniquely correct conformation of a given protein molecule. There is increasing evidence that proteins structurally altered by the incorporation of canavanine instead of arginine do not function properly. Herein lies an important toxic effect of canavanine on most insects.

My colleague Douglas L. Dahlman and I have established experimentally the potent insecticidal properties of canavanine. In our work we employed the tobacco hornworm, *Manduca sexta,* a lepidopteran consumer of plants such as tobacco and the tomato, none of which store canavanine. Hence *M. sexta* has lacked the opportunity to adapt to this toxic amino acid and nullify it. When we injected canavanine labeled with radioactive carbon (C-14) into the hemolymph, or circulatory fluid, of the tobacco hornworm, we found that at least 3.5 percent of the labeled canavanine was incorporated into newly formed insect proteins.

Providing the fat body, an internal organ, of the migratory locust *Locusta migratoria migratorioides* with canavanine results in its incorporation into vitellogenin, a major protein of the fat body that is critically important in egg development. The canavanine-containing vitellogenin of the locust exhibits a higher mobility in an electric field than the natural protein does. This finding reveals a significant change in the physicochemical properties of the protein resulting

from the incorporation of canavanine. The canavanine is assimilated exclusively at the expense of the protein's content of arginine.

In 1973 I learned of the work of Daniel H. Janzen, a tropical ecologist who is now at the University of Pennsylvania. Part of Janzen's work in Costa Rica involved studies of the interaction of the bruchid beetle *Caryedes brasiliensis* and the vinelike legume *Dioclea megacarpa*. My interest in this particular insect-plant interaction developed from Janzen's report that the seeds of the plant contain a considerable amount of canavanine. (Later I determined that canavanine can account for as much as 13 percent of the dry weight of the seed, exclusive of the seed coat, and for as much as 55 percent of all the nitrogen in the seed.) The accumulation of so much canavanine provides a highly effective chemical barrier to predation: the only known insect consumers of the seeds are the larvae of *C. brasiliensis*.

The female beetle lays her eggs on the wall of the newly ripe fruit of the plant, often near the fissure that is characteristic of legumes. The eggs are protected in structures termed oothecae. The time of laying is late fall, at the end of the rainy season and about a month before the pods mature. *D. megacarpa* pods are suitable for the deposition of eggs for only a few months. Bruchid beetles do not lay their eggs on mature or rotting pods, and if the females miss the right period, the beetle population must await the end of the next rainy season before the life cycle is completed with the deposition of eggs on the next crop of pods.

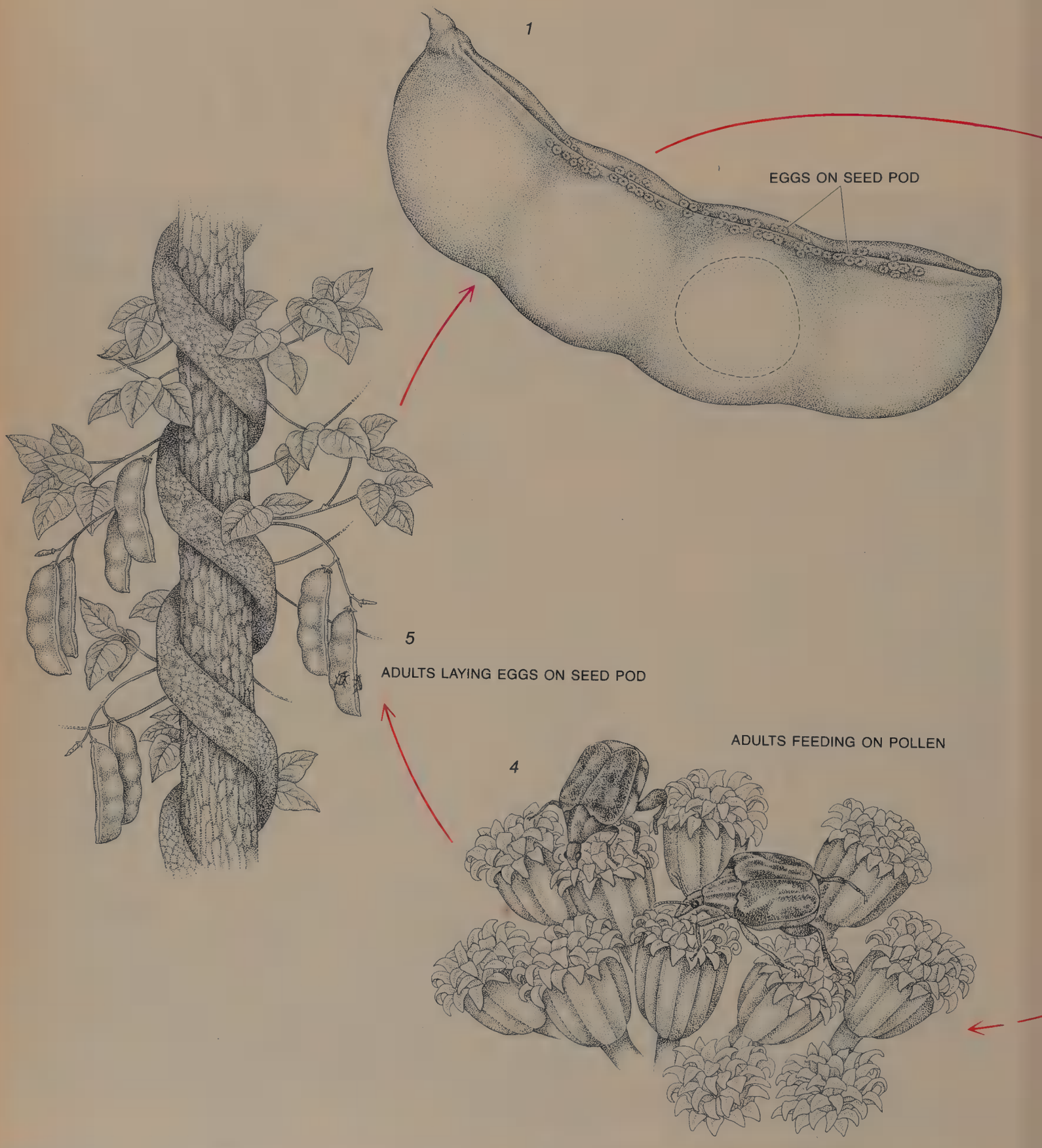
The newly hatched larva bores through the wall of the fruit and the seed coat and makes a small chamber in the storage tissues of the seed. The chamber is the larval home for the several months required for the larva to grow, pupate and emerge as an adult. As many as 50 of the beetles can develop in a single



ADULT BEETLE of the family Bruchidae, the seed beetles, emerges from a seed of the legume *Dioclea megacarpa* after spending several months in it during the larval and pupal stages. The beetle (*Caryedes brasiliensis*) has cut a hole in the coat of the seed in order to get out. The plant makes the amino acid canavanine, which is toxic to other insects. *C. brasiliensis* is its only insect predator. Through a series of biochemical adaptations the larvae (as many as 50 of them in a single seed) avoid the toxic effect and utilize the nitrogen of canavanine.

seed. The young beetle carves an exit portal in the seed coat and escapes ultimately through the suture along the pod or through an opening created by a vertebrate that tears apart the pod to gain access to the seeds. The adult beetle, released from April through June, feeds solely on pollen and does not ingest canavanine.

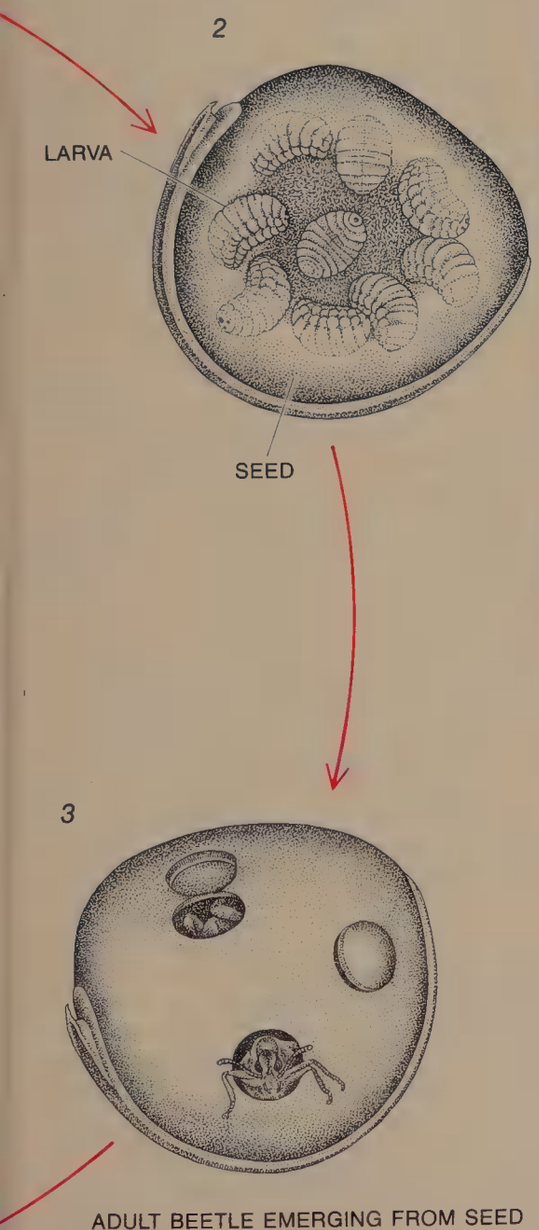
This seed predator is an ideal system for investigating questions of how insects adapt to the toxicants produced by plants. The ability of the beetle to pursue its specialized feeding habit so successfully results from biochemical adaptations that enable it to cope with the canavanine, which would otherwise be highly poisonous. As part of a research program in which I was sponsored by the National Science Foundation I undertook a long-term, systematic investi-



LIFE CYCLE of a new generation of the bruchid beetle *C. brasiliensis* begins when the adult female lays her eggs on a pod of the vinelike legume *D. megacarpa*. The eggs are laid in the late fall, about a month before the pods mature. When the larvae emerge, they bore through the wall of the pod and the coat of a seed, taking up residence in the seed tissues. After several months the larvae have metamorphosed into pupae and then into adults; the young beetles cut exit portals in the seed coat, emerging into the pod and thence through the suture of

gation of these biochemical adaptations.

As a beginning my associates and I asked whether or not this bruchid beetle incorporates canavanine into its proteins. If it avoids producing aberrant proteins, how does it manage to do so? To answer these questions we injected



the pod or an opening made by an animal other than an insect. The adult beetles of the species feed on the pollen of various plants, beginning the cycle again by laying eggs in the fall.



For personally signed Ken Davies print, 18" x 19", send $10. payable to "ANCO", Box 2817-SN, NYC, 10163

Why It's Such A Rare Bird

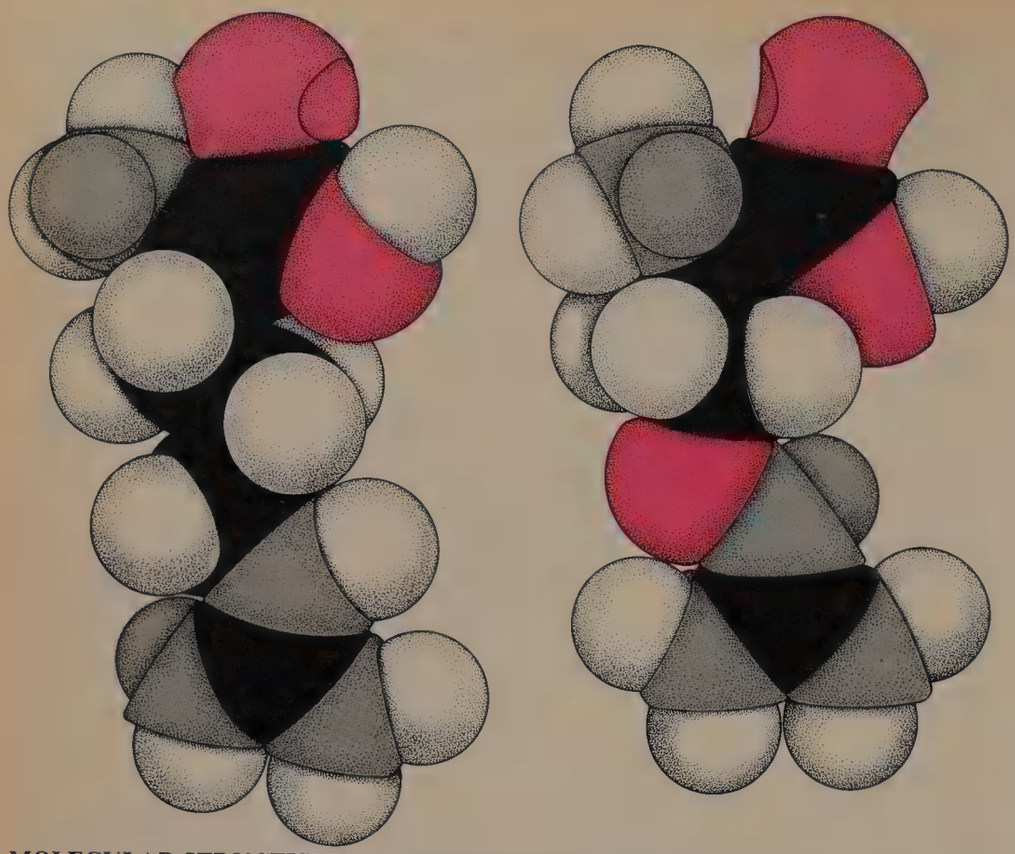
Wild Turkeys are masters of camouflage and evasion. A large flock of birds will lie quietly within yards of a man passing through the forest, and never be seen.

The Wild Turkey is truly a native bird, unique to America. And it is the unique symbol of the greatest native whiskey in America–Wild Turkey.

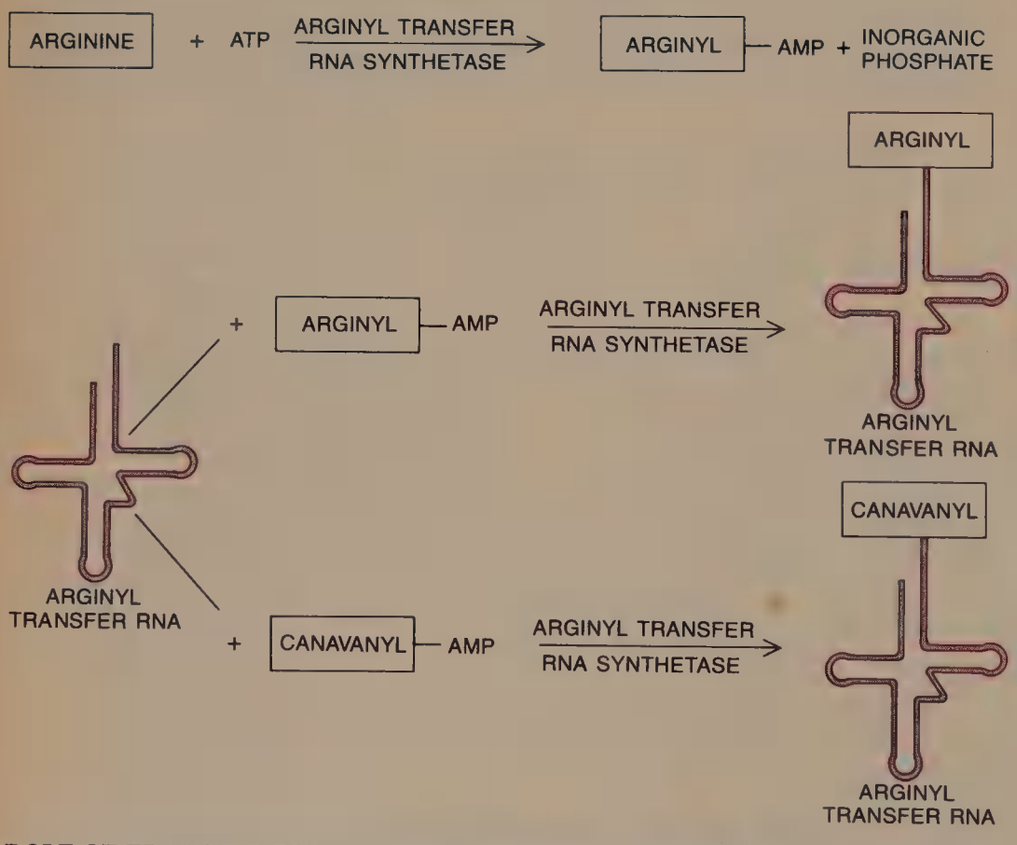


WILD TURKEY® / 101 PROOF / 8 YEARS OLD

Austin, Nichols Distilling Co., Lawrenceburg, Kentucky © 1981



MOLECULAR STRUCTURES of L-arginine (*left*) and L-canavanine (*right*) are quite similar. The constituent atoms are carbon (*black*), hydrogen (*white*), nitrogen (*gray*) and oxygen (*color*). Arginine is a protein-forming amino acid; canavanine is one of some 250 amino acids synthesized by higher plants but not utilized to make proteins. The structural difference between the two amino acids is small: the terminal methylene group ($-CH_2$) of arginine is replaced by oxygen in canavanine. In addition the guanidino group of arginine is in the imino form ($=NH$), whereas the corresponding group in the canavanine molecule is in the amino form ($-NH_2$).



ROLE OF TRANSFER RNA in dealing with arginine or mistakenly with canavanine in the production of proteins is shown. Normally (*top*) the enzyme arginyl transfer RNA synthetase reacts with arginine in the presence of adenosine triphosphate (ATP) to make an activated form of the amino acid. The enzyme also mediates the linkage of activated arginine to the correct transfer RNA. Usually (*middle*) arginyl transfer RNA synthetase binds arginine to its transfer RNA before delivering arginine to a ribosome complex, where protein is made. If canavanine is bound instead of arginine (*bottom*), it is transported to the assembly site by the transfer RNA molecule that normally carries arginine. Protein synthesis does not require verification of the amino acid attached to the transfer RNA, and so canavanine ends up in the protein.

larvae, obtained from infected *D. megacarpa* seeds, with radioactively labeled canavanine. Only the terminal carbon atom of the injected canavanine was radioactive. This unique labeling pattern was important because the treatment of radioactive canavanine with the enzymes arginase and urease releases the radioactive carbon atom as carbon dioxide, a gas that can be trapped chemically and then quantified by liquid-scintillation spectroscopy.

The newly synthesized and slightly radioactive proteins obtained from the larvae were isolated and digested with a strong acid to release their constituent amino acid building blocks. The protein digest was then purified by ion-exchange chromatography, which isolated basic amino acids, such as canavanine. Treatment of the isolated amino acids with arginase and urease failed to cause the release of appreciable amounts of radioactive carbon dioxide. This finding established that the larvae were not incorporating canavanine into protein in significant amounts, although some insertion of the radioactive carbon atom of canavanine into protein amino acids had occurred.

The biochemical basis for the finding became apparent when the arginyl transfer RNA synthetase of this insect was compared with that of *Manduca sexta,* the tobacco hornworm. The enzyme from *M. sexta* attached both arginine and canavanine to arginyl transfer RNA, but the arginine-activating enzyme from *C. brasiliensis* attached only arginine. It thus appeared likely that among the biochemical adaptations achieved by the beetle was the development of an arginyl transfer RNA synthetase capable of discriminating between canavanine and arginine. In this way canavanine is not linked to the transfer RNA for arginine and the beetle avoids the production of aberrant, canavanine-containing proteins.

In a recent companion study we gained additional insight into the discriminatory capability of the protein-synthesizing system of the bruchid beetle. The study required the synthesis of a group of radioactively labeled amino acids related structurally to arginine and their testing, along with canavanine and arginine, to determine if any of them were incorporated into the protein of *M. sexta* and *C. brasiliensis.* The tobacco-hornworm larvae fixed each of the compounds into newly synthesized protein; the bruchid-beetle larvae did not. The marked ability of this seed predator to distinguish arginine from molecules structurally akin to it confers a general resistance to the error of incorporating the wrong amino acids into proteins.

Many important questions remained, but the one we directed our attention to was how the bruchid-beetle larva han-

dles canavanine. It may excrete the substance or in some other way avoid its adverse biochemical effects. An exciting alternative possibility was that the beetle had developed some means of utilizing the toxic natural chemical as a food resource. It was reasonable to think the insect might not sacrifice such an abundant source of nitrogen.

Our first effort was to analyze the canavanine content of *D. megacarpa* seeds infested with bruchid-beetle larvae. We applied a statistical technique to estimate the weight of the seed prior to insect attack from the weight of the intact seed coat. The amount of canavanine in the uneaten part of the seed and in the frass (the fecal matter of the larvae) enabled us to calculate how much canavanine the seed had contained originally. These calculations indicated that more than half of the original seed canavanine was consumed in larval feeding.

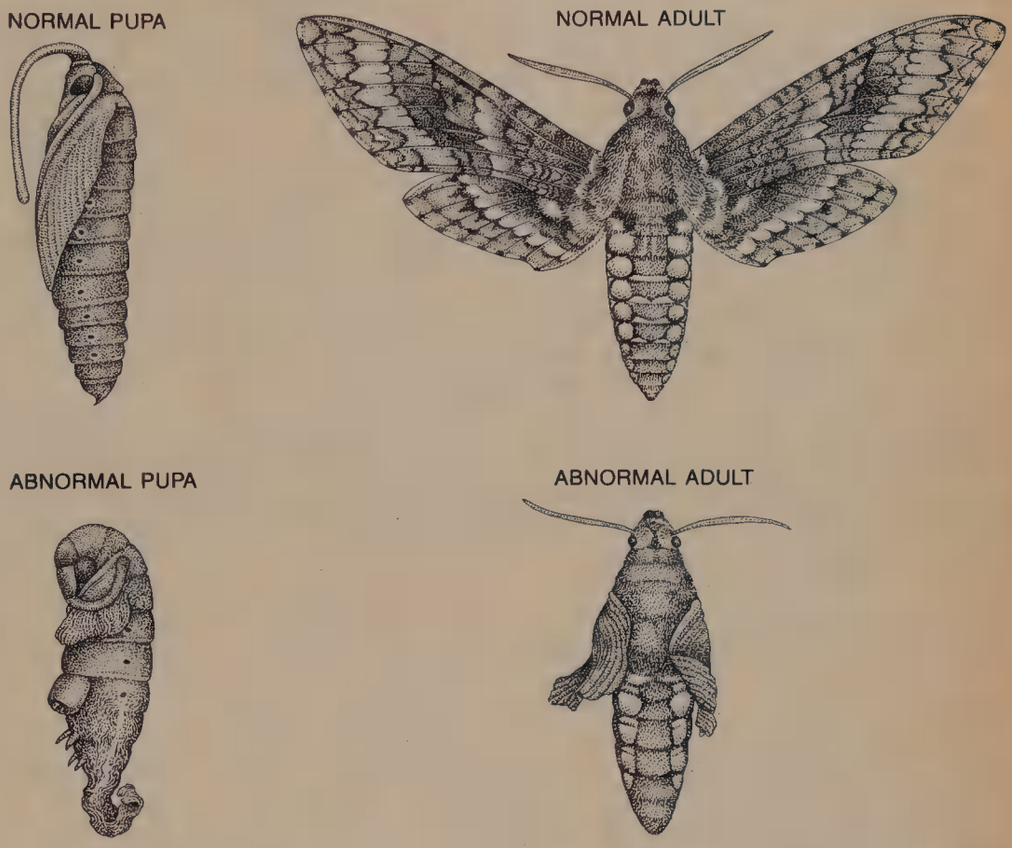
This finding still left the question of how the ingested canavanine was metabolized. Arginase, an enzyme that cleaves L-arginine into L-ornithine and urea, also acts on L-canavanine, cleaving it into L-canaline and urea. The enzyme is distributed widely among insects and has several important functions, one of them being to provide ornithine from arginine for the formation of glutamic acid. The bruchid beetle produces arginase and would be able to convert canavanine into canaline and urea: $H_2N(C{-}NH_2){=}N{-}O{-}CH_2{-}CH_2{-}CH(NH_2)COOH$ [L-canavanine] $\longrightarrow$ $H_2N{-}O{-}CH_2{-}CH_2{-}CH(NH_2)COOH$ [L-canaline] + $H_2N{-}C(=O)NH_2$ [urea].

Urease, the enzyme that acts on urea, is seldom found in insects. Our analysis of the bruchid-beetle larvae disclosed an extraordinarily high urease activity. The larvae cleave canavanine into canaline and urea and then draw on urease to generate ammonia and urea. In this way half of the nitrogen stored in canavanine becomes available as ammonia for metabolic reactions.

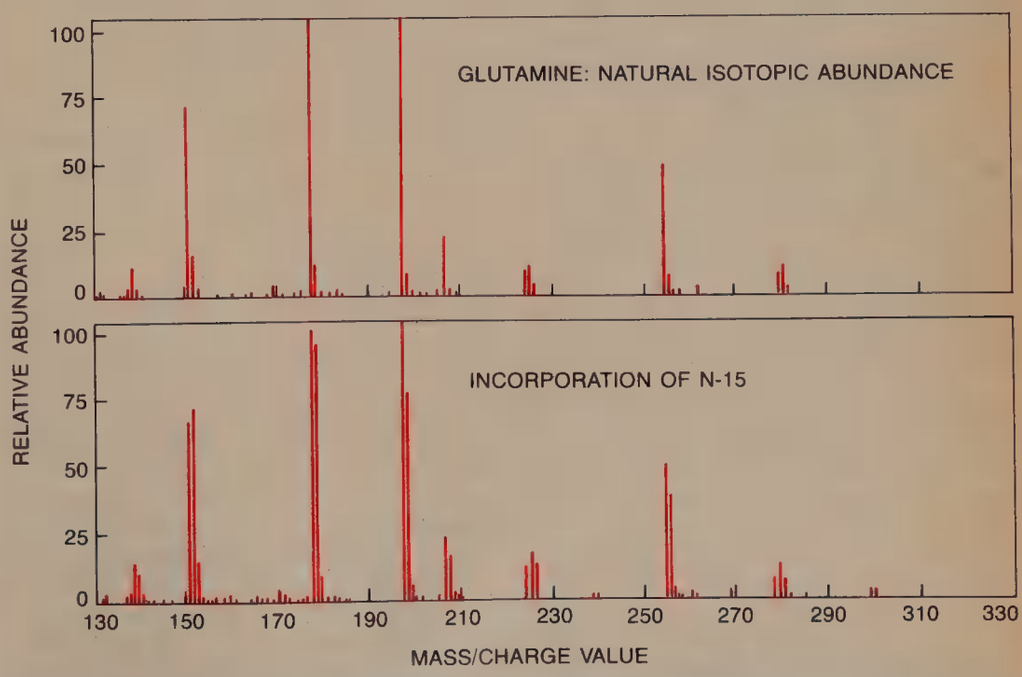
Does the larva utilize the ammonia as a source of nitrogen for its metabolic reactions? Is it not only detoxifying canavanine but also incorporating part of the nitrogen from it into newly made amino acids? These are important questions because with the possible exception of the incorporation of cyanide into L-asparagine, an amino acid of proteins, no other instance is known of an insect utilizing a toxic plant compound for amino acid production.

The commercial availability of urea labeled with the heavy isotope of nitrogen (N-15) and the facile ability of the bruchid-beetle larva to convert canavanine into urea made it possible to test the hypothesis experimentally. Larvae were injected with urea containing the heavy nitrogen isotope, and the newly synthesized free amino acids were isolated and converted chemically into a more volatile form. This conversion made it possible to separate the amino acids by gas chromatography. As each compound emerged from the chromatographic apparatus it entered a mass spectrometer. (These experiments were done at the excellent mass-spectroscopy facilities of the Tobacco and Health Research Institute of the University of Kentucky.)

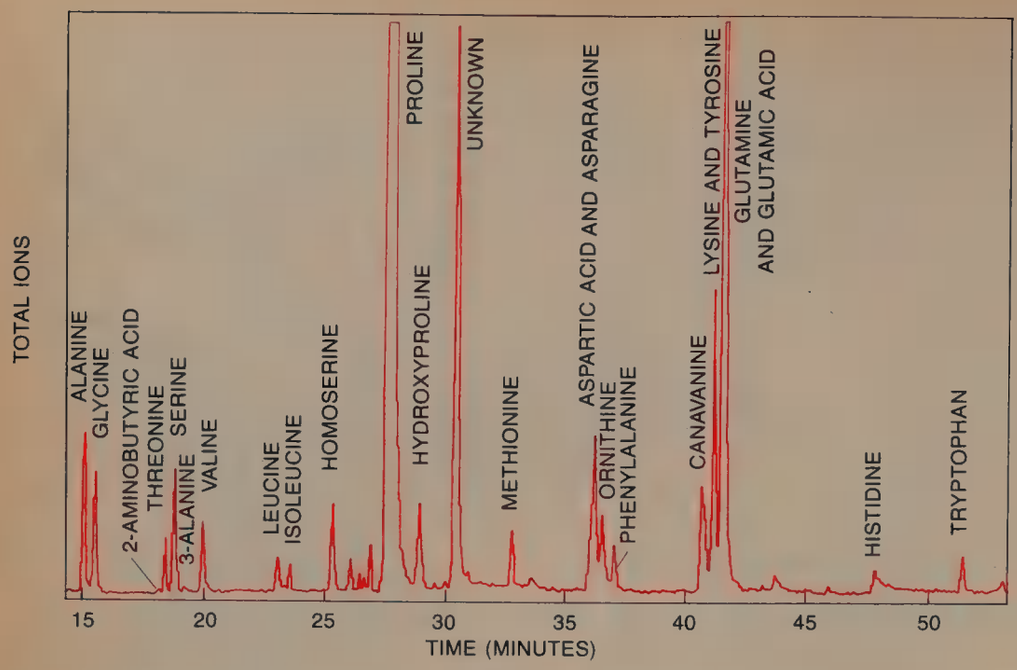
In the mass spectrometer the compound is bombarded by electrons that break the amino acid into a unique pat-



EFFECTS OF CANAVANINE are seen in *Manduca sexta*, the tobacco hornworm. The abnormal pupal and adult forms resulted from the inclusion of canavanine in the diet fed to the larvae in the laboratory many days earlier. Comparably severe developmental abnormalities are elicited by injecting canavanine directly into the hemolymph, the circulatory fluid of the insect.

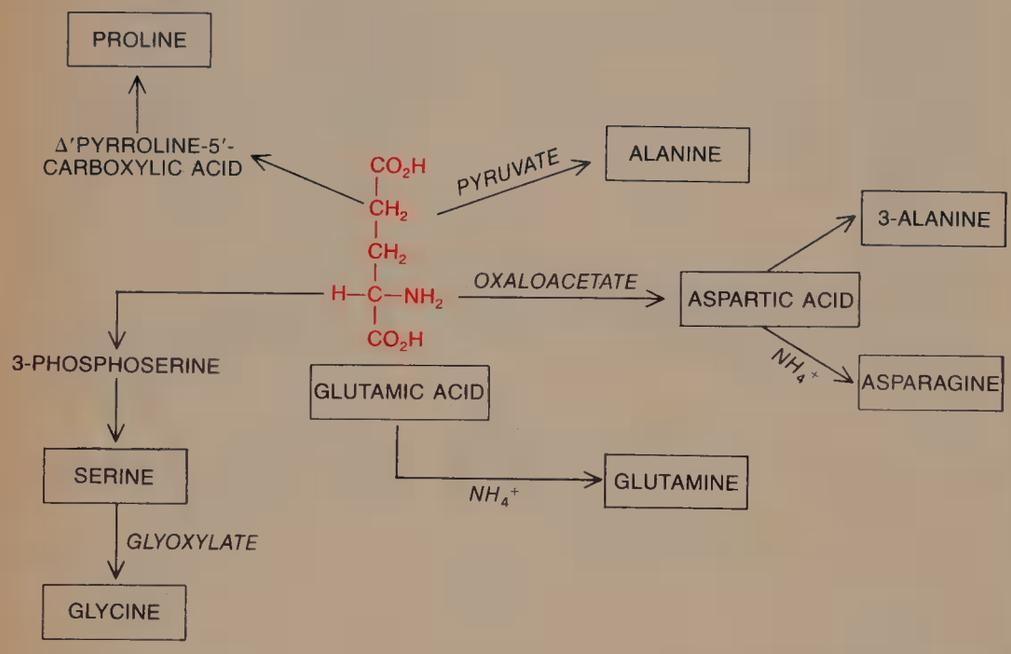


INCORPORATION OF NITROGEN from canavanine into amino acids was determined by mass spectroscopy. Larvae of *C. brasiliensis* were injected with urea labeled with the heavy isotope of nitrogen (N-15). The amino acids they synthesized were separated by gas chromatography, and each compound then entered the mass spectrometer. The results show that the larvae incorporated N-15 into the protein amino acids that they can synthesize but not into the ones that they must secure in their diet. One of the protein amino acids that incorporated significant amounts of N-15 was L-glutamine; the mass-spectrometric data for it are represented here.

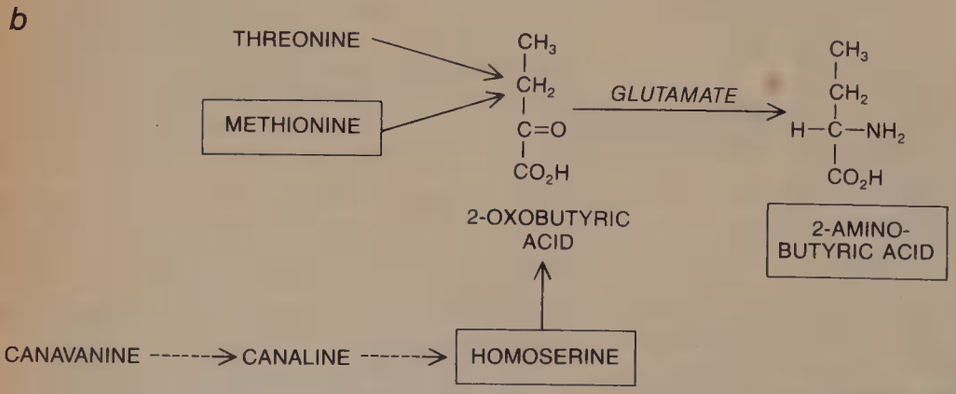


COMPUTER ANALYSIS of the incorporation of heavy nitrogen into the amino acids of *C. brasiliensis* yielded these results. The data show that the nitrogen of canavanine, released by the formation and degradation of urea, is utilized by the bruchid beetle in the synthesis of the amino acids it can manufacture but not in the amino acids the insect can get in its normal diet.

a



b



POSSIBLE BIOSYNTHETIC PATHWAYS from glutamic acid (*color*) to the N-15-labeled amino acids produced by the larva of *C. brasiliensis* are charted (*a*). The amino acids that took up heavy nitrogen are indicated by rectangles. The formation of 2-aminobutyric acid from methionine and homoserine, both of which were found to contain N-15, is also shown (*b*). The synthesis of homoserine by *C. brasiliensis,* the only insect known to do so, cannot have been through the degradation of canavanine by way of canaline, because the canavanine was not labeled.

tern of fragments. Computer analyses yield an accurate characterization and compilation of the fragments obtained for each compound. Knowing the relative abundance of the fragments containing the two isotopes of nitrogen makes it possible to accurately reconstruct the proportions of N-15 and N-14 in the original amino acid.

These determinations revealed a significant incorporation of heavy nitrogen into alanine, glycine, serine, proline, methionine, aspartic acid (and/or asparagine) and glutamic acid (and/or glutamine). They are the protein-forming amino acids that insects can synthesize from suitable precursors. On the other hand, no appreciable heavy nitrogen was found in threonine, leucine, isoleucine, histidine, lysine or hydroxyproline. They are the protein-forming amino acids that insects cannot synthesize and must get from their food. Valine is also taken to be essential in the insect diet, but *C. brasiliensis* was able to synthesize very small amounts of it. Tyrosine, phenylalanine, arginine and tryptophan were not evaluated because of technical limitations. Interestingly enough, the bruchid-beetle larvae also incorporated appreciable amounts of N-15 into several nonprotein amino acids: 2-aminobutyric acid, 3-alanine and homoserine.

How might the nitrogen of canavanine that is ultimately released as ammonia support the manufacture of these amino acids? It may be that glutamic acid and glutamine play a central role in the amino acid metabolism of the insect. Enzymes that can catalyze the synthesis of alanine and aspartic acid from glutamic acid are distributed widely among insects. Glutamine and asparagine are readily obtained from glutamic acid and aspartic acid, respectively, by the simple addition of ammonia to those amino acids. Glutamic acid is also linked directly to the synthesis of proline and serine; glycine forms readily from serine. The considerable amount of heavy nitrogen found in 2-aminobutyric acid may seem unexpected, but the compound forms from an organic acid that is readily synthesized from threonine and methionine. Thus virtually all the amino acids known to be synthesized by insects, including those synthesized by *C. brasiliensis,* contain appreciable amounts of nitrogen equivalent to the amount readily supplied by canavanine.

A further problem is that L-canaline, a nonprotein amino acid produced by the cleavage of L-canavanine into urea, is highly toxic in its own right. It is distinctive in being the only nonprotein amino acid with a free terminal aminooxy group: $-ONH_2$. Experiments with larvae of the tobacco hornworm revealed that L-canaline retards growth, causes severe developmental aberrations, increases mortality and interferes with nerve function. Canaline reacts with the

aldehyde group of pyridoxal phosphate, a cofactor essential for the function of certain enzymes. The canaline–pyridoxal-phosphate complex is stable, and its formation essentially shuts down the catalytic action of certain enzymes that contain pyridoxal phosphate. On first consideration it appears that the bruchid beetle is merely exchanging one poison for another.

By examining the metabolic fate of canaline in the bruchid-beetle larvae we found that the insect has an enzyme capable of cleaving canaline to yield homoserine: $H_2N–O–CH_2–CH_2–CH(NH_2)COOH$ [L-canaline] $\longrightarrow$ $HO–CH_2–CH_2–CH(NH_2)–COOH$ [L-homoserine] + NH_3. With this finding we confronted the question of how the insect deals with all the ammonia, which is a toxic form of reduced nitrogen. Uric acid is the principal nitrogenous excretory product of insects; ammonia is normally a minor excretory product. Analyzing the frass of the larvae, we found that nearly 90 percent of the nitrogen in it was in the form of ammonia and urea. Only 11 percent of the total excreted nitrogen was in the form of uric acid. Clearly this seed-eating bruchid beetle can excrete some of the extraneous ammonia directly in the frass.

The elimination of ammonia by way of the frass is not the only mechanism available to the insect for avoiding the toxic effects of too much ammonia. The larva of the beetle has a highly active glutamine synthetase, an enzyme that can add ammonia to glutamic acid to form glutamine. Glutamine is a principal free amino acid of the insect's hemolymph. Analysis of the amino acids of the hemolymph also reveals a massive peak for the amino acid proline. Perhaps proline serves as a source of carbon skeleton for the formation of glutamic acid before the formation of glutamine. The reaction sequence from proline to glutamic acid to glutamine may provide an important means of channeling toxic ammonia into a harmless form.

As for homoserine, the investigations of John Giovanelli and his associates at the National Institute of Mental Health have established that plants employ homoserine derivatives as precursors in the manufacture of certain indispensable amino acids containing sulfur. One would expect many types of insect (for example, the aphids that feed on the seed of the pea plant, *Pisum sativum*) to utilize homoserine because this nonprotein amino acid can account for more than 10 percent of the dry weight of certain parts of a plant.

After administering homoserine labeled with carbon 14 to the bruchid-beetle larvae we found a detectable but quantitatively insignificant transfer of the radioactive atoms to the amino acids methionine and cysteine. Automated amino acid analyses of the extract of these homoserine-treated insects revealed a massive peak of radioactivity that was not caused by an amino acid. This was evident because the unknown substance did not form a colored complex with ninhydrin. That reagent reacts with amino acids to form typically a blue purple color; it is the classic colorimetric test for amino acids.

It may be that the bruchid beetle uses homoserine to provide nitrogen for the synthesis of other amino acids through the action of an aminotransferase. The reaction of this enzyme with homoserine would cause the homoserine to lose its normal ninhydrin-positive response. In any event it is clear that the insect is synthesizing significant amounts of homoserine for a purpose other than the manufacture of amino acids containing sulfur.

A fascinating aspect of these biochemical processes is that the bruchid beetle has adapted to a detrimental substance in its food (canavanine) by functioning in much the same way as a plant synthesizing the same substance. This convergent evolution is revealed by four lines of evidence. The first is the formation of an enzyme (arginyl transfer RNA synthetase) capable of discriminating between arginine and canavanine. The ability to distinguish between an amino acid that is normally part of a protein and one that is not is a well-recognized mechanism whereby plants avoid the adverse effects of their own nonprotein amino acids.

The second line of evidence is that the formation of urea from canavanine coupled with the formation of ammonia from urea is the principal means by which plants that make canavanine mobilize the nitrogen stored in it. The third is that the insect detoxifies canaline by converting it into homoserine; in plants synthesizing canavanine the conversion of homoserine into canaline may be part of the synthesis of canavanine.

Finally, we found that the arginase of *C. brasiliensis* has a greater apparent affinity for canavanine and reacts faster with canavanine (compared with arginine) as a substrate than the arginases of the other insects we tested. A similar distinction is found in comparing the arginase of *Canavalia ensiformis*, a legume that utilizes canavanine, with the arginase of the soybean *Glycine max*, which does not synthesize canavanine.

Another conclusion to be drawn from our investigations is that this seed-eating bruchid beetle has achieved many distinct but interrelated biochemical adaptations in its utilization of *D. megacarpa* as a food source. It is reasonable to propose that among the first of the adaptations was the development of the discriminatory ability enabling the insect to avoid the synthesis of dysfunctional proteins. The bruchid beetles that originally invaded the legume may have simply excreted canavanine and canaline or somehow avoided them. Other nitrogen-rich metabolites stored in the seed (such as proteins) could have satisfied the need of the growing larva for nitrogen. Over a period of time, through natural selection, the insects may have become progressively better equipped to cope with canavanine and canaline and ultimately could utilize the nitrogen they contain.

Such adaptive successes would have fostered the dependence of the beetle on *D. megacarpa*. In time the benefits of its association with the plant may have led to the failure to lay its eggs on other seeds in its habitat. In considering the narrow feeding range of *C. brasiliensis* and the danger inherent in depending on a single food and egg-laying resource, it is important to think of the potential advantages. The insect is relatively free from competition for the seeds and does not have to invest in a capacity to process a wide range of plant toxins. The developing larvae are assured a reasonably safe haven, since the toxicity of canavanine should protect the seed from the larvae of other species.

A final point on which these studies bear is the remarkable ability of insects to adapt to a host of detrimental substances, including the insecticides devised by human beings. Is it possible that microbes living symbiotically in the insect gut have a major role in processing toxic compounds such as canavanine? The seed of *D. megacarpa* has an array of microbial symbionts and pathogens that *C. brasiliensis* could have acquired, and they may account for some or even all of the biochemical adaptations required for the utilization of canavanine and canaline. We have isolated microbial symbionts from the beetle larvae and found several that can subsist on canavanine and canaline (or one of them) as the sole source of both carbon and nitrogen. I intend to ascertain how much of the acquired biochemical abilities of this seed predator derives from its microbial symbionts.

Our continuing studies of the relation between the bruchid beetle and its leguminous host have provided a basic understanding of how an insect has adapted at the biochemical level to toxins presented by its host plant. It is regrettable that similar biochemical investigations have not been done, so that the present information is limited to this single association. Nevertheless, what we have found is of practical value in helping to understand the response of insects not only to natural toxins but also to chemical control measures. In addition the studies reveal the kinds of insight that can be gained by the union of ecological and biochemical approaches.

The Eruption of Krakatau

The explosions that obliterated most of the Indonesian island 100 years ago are only now beginning to be understood. The evidence is largely the volcanic deposits and the timing of air and sea waves

by Peter Francis and Stephen Self

One hundred years ago, on the morning of August 27, 1883, a series of intermittent volcanic explosions culminated in the paroxysmal blasts that accompanied the destruction of most of Krakatau, a small island in the Sunda Straits. Tsunamis ("tidal waves") set in motion by the eruption killed more than 30,000 people on the neighboring Indonesian islands of Java and Sumatra. The event attracted worldwide attention, and the ensuing scientific investigations made important contributions to the fledgling field of volcanology. Only now, however, is it becoming possible to explain some of the major events in the eruption sequence in terms of the underlying volcanic processes. Here we address three much-debated questions the eruption raises: What triggered it? Why were there so many violent explosions in the eruption sequence? What was the relation between the devastating tsunamis and the massive explosions?

When Krakatau erupted, the explosions were heard in central Australia, Manila, Sri Lanka and on Rodriguez Island, more than 5,000 kilometers away in the Indian Ocean. Lower-frequency atmospheric waves (at frequencies too low to be audible) were detected worldwide; barometers in Tokyo, 5,863 kilometers away, registered a pressure increase of 1.45 millibars. The sea waves generated by the eruption traveled not only across the Pacific but also across the Atlantic: they were detected by tide gauges in the Bay of Biscay, 17,000 kilometers away. The dust and gases injected into the atmosphere by the eruption produced spectacular sunsets worldwide for months afterward. The mean temperatures recorded in the Northern Hemisphere during the same period were from .5 to .8 degree Celsius lower than normal.

Krakatau's reputation as the classic volcanic eruption, however, is perhaps due as much to the date of its occurrence as to its violence. The eruption, which was one of the first to be the subject of intensive scientific investigation, came at a time in the Victorian era when science was followed by a large and enthusiastic audience. The eruption of the volcano Tambora on the Indonesian island of Sumbawa in 1815 attracted comparatively little attention even though it was much larger; Tambora is estimated to have ejected between 150 and 180 cubic kilometers of pumice and ash whereas Krakatau ejected 20 cubic kilometers. Moreover, the eruption of Tambora caused the death of more than 90,000 people either directly or as a result of tsunamis and an ensuing famine [see "The Year without a Summer," by Henry Stommel and Elizabeth Stommel; SCIENTIFIC AMERICAN, June, 1979]. At the time, however, no one made the connection between the unusually cold weather in Europe and North America in the summer of 1816 and the eruption of Tambora the year before, and the eruption has yet to be studied in detail.

In contrast, both the Royal Society of London and the Dutch government, which was then the colonial administrator of the Indonesian islands, published lengthy reports about the eruption of Krakatau soon after it occurred. The Royal Society's report emphasized the worldwide atmospheric effects of the eruption. Of its 494 pages 312 were devoted to "the unusual optical phenomena of the atmosphere, 1883–6, including twilight effects, coronal appearances, sky haze, coloured suns and moons, etc." The report issued by the Dutch committee, led by Rogier D. M. Verbeek, a mining engineer and geologist, covers the geologic aspects of the eruption in greater detail. The members of the committee visited the scene of the eruption on October 15, 1883, and repeatedly thereafter. They mapped the new islands and the remnants of the old ones and measured the changed contours of the ocean bottom. Verbeek himself collected samples of the volcanic ejecta, which he and his team later examined under the microscope.

The conclusions Verbeek drew from this diverse evidence were for the most part remarkably prescient. For example, on the basis of depth soundings and measurements of the area of ashfall he made an estimate of the amount of material ejected by the volcano that has stood without substantial revision to this day. Finding that the rock specimens he had collected were magmatic rock rather than older rock, he correctly proposed that the old volcanic cone had not been blasted into the air but had foundered into the sea when the magma chamber underlying it was exhausted.

The full story of the eruption based on the accounts of witnesses makes compelling reading. Here we shall concentrate on those events of the eruption sequence that seem to be correlated with the emplacement of volcanic deposits, because it is on this correlation that reconstructions of what happened must be based.

The eruption sequence has been established largely from records kept by Dutch administrators living well above the shoreline in Sumatra and Java and in towns in the interior and from reports made by the officers on watch on ships passing through the Sunda Straits; several of the ships sailed near Krakatau during the most violent part of the eruption. The descriptions of the nature and stratigraphy of the volcanic deposits are based largely on a field study by one of us (Self) and Michael R. Rampino of the Goddard Institute for Space Studies of the National Aeronautics and Space Administration. We studied and sampled the deposits when we visited the island in 1979 in the course of a broader investigation into the atmospheric effects of volcanic eruptions.

Navigational charts of Krakatau and the surrounding areas of the Sunda Straits made before the eruption show that the island consisted of three volcanic cones aligned roughly northwest to southeast. The largest cone, Rakata, 813 meters high, was at the southern end of the chain. A lower cone, Danan, was in



BISECTED VOLCANIC CONE is all that remained of the island of Krakatau after the eruption of August 27, 1883. The island originally consisted of three volcanic cones aligned roughly northwest to southeast. The main vent of the 1883 eruption is thought to have lain between the two northernmost cones. Late in the eruption sequence two-thirds of the island foundered into the sea as the roof of the magma chamber underlying it collapsed. The northern face of the southern cone, Rakata, which was perched on the edge of the new caldera (the submarine depression formed by the collapse), was left virtually unsupported and subsequently slumped into the sea. Clearly visible are the interior structure of the old volcanic cone (including the rock whitened by hydrothermal alteration near the central vent), the feeder dikes (dark columns) leading down to the magma chamber and the alternating lava flows and ash layers that made up the cone. The layer of white pumice deposits laid down on both flanks of the cone is also visible in this chromolithograph from *Album of Krakatau.*



TWO NEW ISLANDS named Steers and Calmeyer in the Sunda Straits north of Krakatau consisted of deposits laid down on the floor of the straits that were exposed above sea level. The islands were created by successive pyroclastic flows: ground-hugging clouds of pumice and ash driven by gravity and fluidized by hot gases. The flows traveled an average of 15 kilometers from the vent, much of it through or over water. Hot material caused the explosive vaporization of seawater; some of the many large explosions late in the eruption sequence may have been such secondary explosions. This chromolithograph from *Album of Krakatau* shows a large secondary-explosion crater on Calmeyer. The crater closely resembles the craters formed by pyroclastic flows from Mount St. Helens that entered Spirit Lake. *Album of Krakatau* was published in 1886; the loosely packed deposits eroded quickly and the islands soon vanished below the surface.

the middle and a much lower one, Perbuwatan, was at the northern end. Lava flows around Perbuwatan showed that it had been active in the geologically recent past, and it is thought to have been the site of a pumice eruption in 1680. Two smaller islands near Krakatau—Sertung and Rakata Kecil (Little Rakata)—and the southern end of Krakatau itself were probably remnants of the rim of a submerged caldera: a large volcanic crater formed by collapse.

When Perbuwatan burst spectacularly into life with a series of deafening explosions on May 20, 1883, after nearly 200 years of inactivity, the eruption came as a surprise. Krakatau was uninhabited and was visited only occasionally by fishermen and woodcutters. Thus any small-scale activity that might have preceded its reawakening went unnoticed. There was a period of markedly increased seismic activity around the straits before the eruption, but at the time no link was made between this activity and Krakatau.

The May eruption of Perbuwatan was accompanied by explosions that could be heard more than 150 kilometers away. At the same distance atmospheric pressure waves of very long wavelength were energetic enough to stop clocks, rattle doors and windows and dislodge hanging lamps. Since the pressure waves were inaudible, their effects were often mistaken for those of earthquakes. Although some seismic activity was recorded during the climactic phase of the eruption in August, at this stage almost all the energy seems to have been transmitted through the air.

Perbuwatan continued to erupt intermittently throughout May, June and July, but the activity was relatively unimpressive. According to Captain Ferzenaar, a Dutch surveyor who visited Krakatau on August 11, trees were still standing, although they had been stripped of foliage by the falling ash. Ferzenaar noted that a layer of ash about 50 centimeters thick covered the island. Today only very limited outcrops of fine-to-medium-grained ash from this early phase of the eruption can be seen.

The relatively minor explosions continued into August and culminated in massive explosions on August 26 and 27. The scale of activity on these two days was so large that it has been difficult to piece together exactly what happened, and much detail is surely missing. No one within close range of the volcano survived. Reconstructions of the eruption sequence in this crucial period are based largely on instrument records and the volcanic deposits. The enormous explosions on August 27 generated air waves so powerful that they registered on a recording pressure gauge at the gasworks at Jakarta, which thereby preserved a record of their timing and relative amplitude. The timing and magnitude of the tsunamis were recorded by tide gauges along the straits.

The climactic phase of the eruption can be divided into two stages on the basis of the nature and sequence of the deposited ash layers. Beginning at 1:00 P.M. on August 26 a series of explosions at intervals of about 10 minutes created a more or less sustained eruption column over the island that is reported to have reached a height of about 25 kilometers (82,000 feet). The explosions produced primarily air-fall material: pumice and ash carried upward in the atmosphere by a column of hot convecting gas. (Both pumice and ash are frothy glassy materials created by the chilling of vesiculating magma. The distinction between them is largely one of size; fragments smaller than two millimeters in diameter are generally called ash.)

Although pumice and ash deposits from this phase of the eruption accumulated to thicknesses of up to 20 meters on the islands of Sertung and Rakata Kecil and ships within 20 kilometers of the volcano reported heavy falls of ash accompanied by large pumice clasts (fragments) up to 10 centimeters in diameter, the heavy ashfall was limited in extent. Only a minor fall of ash was reported on Sumatra and on western Java.

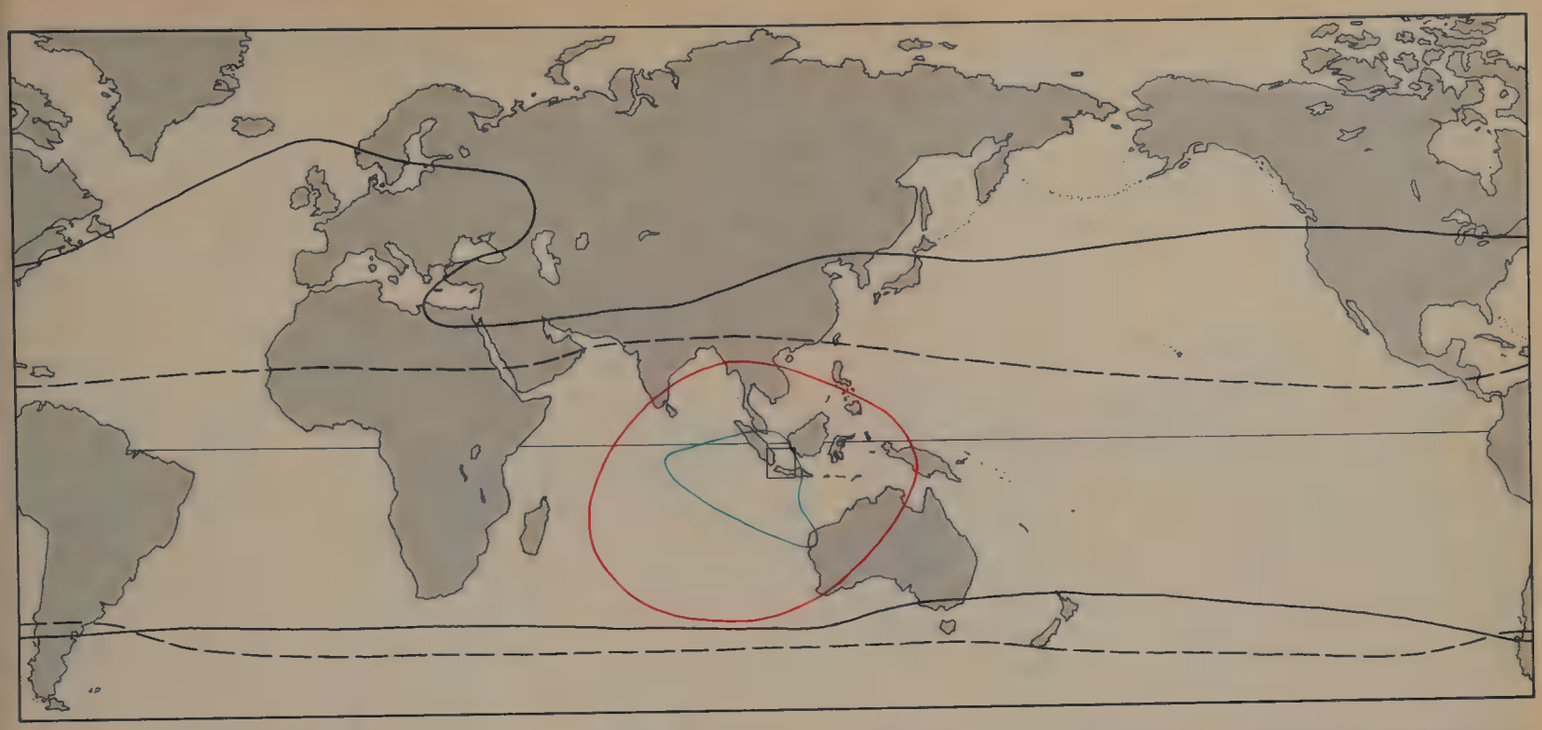
At 5:30 A.M. on August 27 the character of the eruption changed dramatically. In the course of the day there were many enormous explosions. The largest explosion, the one heard as far away as Rodriguez Island, came at 9:58 A.M.; it was associated with the largest tsunami, estimated to have crested at 40 meters, which caused most of the deaths on the neighboring coastlines.

During this period the explosions were characterized by the formation of pyroclastic flows as well as a sustained column of airborne material. The explosions of August 27 seem to have been paroxysmal rather than continuous. Each blast propelled a large amount of pumice and ash high into the air, perhaps as high as five kilometers. The aggregation of material was too dense and heavy to remain airborne for long, however, and most of it immediately fell back to the ground. There it formed incandescent, ground-hugging clouds, driven by gravity and fluidized by hot gases, that moved rapidly off the island and into the sea. A fraction of the ash was lofted much higher by convection currents set up in the atmosphere by local heating. The resulting ash cloud may have reached a height of about 40 kilometers. The ash from the towering cloud fell out over a wide area. The neighboring coastlines were plunged into darkness and ashfall was reported as far away as the Cocos Islands, 1,850 kilometers from Krakatau.

The pyroclastic flows laid down a dis-



KRAKATAU WAS A SMALL UNINHABITED ISLAND about 32 kilometers west of the narrowest part of the Sunda Straits between the Indonesian islands of Sumatra and Java. More than 30,000 people were killed by the eruption that obliterated most of the island. The majority were victims of the tsunamis that swept over the low-lying coastlines of the neighboring islands. Along the east side of Lampong Bay on the southern coast of Sumatra, however, some bodies were found buried in ash. Those people were probably killed by pyroclastic flows that had traveled more than 40 kilometers from Krakatau on the surface of the sea and yet were still lethally hot. The names on the map are locations mentioned in the graph of the eruption sequence on page 152. Modern names have been substituted for their 19th-century equivalents.



GLOBAL EFFECTS of the Krakatau eruption are plotted on this world map. Shown are the range within which ash fell out (*blue line*), the range within which the explosions were heard (*red line*) and the range within which atmospheric effects caused by the volcanic ash and aerosols injected into the upper atmosphere were reported before September 22, 1883 (*broken black line*), and by late in November (*solid black line*). Why the final explosions were so violent has not yet been explained. Ash from the eruption fell out over an area of 700,000 square kilometers, the largest range of ashfall ever produced by a volcanic eruption in historic times. The major explosions generated dense clouds of fine ash over the Sunda Straits that probably attenuated sound quite effectively. One of the consequences may have been that people living on the coastlines of Java and Sumatra, soon to be victims of the tsunamis generated by the eruption, did not hear the explosions that were heard by others as far away as central Australia. A study by the British physicist Rollo Russell of the path and travel time of the high cloud, based on reports of atmospheric effects, provided the first evidence for stratospheric circulation patterns.

tinctive deposit called ignimbrite; these deposits account for the largest fraction of the material erupted by Krakatau. Because the material in the pyroclastic flows was fluidized (given a low density and viscosity) by hot gases, it made extremely efficient use of the kinetic energy it acquired in falling from the height of the eruption column and so traveled long horizontal distances. Deposits of ignimbrite as thick as 40 meters were laid down as far as 15 kilometers from the primary vent, thought to have been between Danan and Perbuwatan.

The pyroclastic flows seem to have spread out preferentially to the north and northeast, covering the islands and the surrounding sea floor with a blanket of ignimbrite. The distribution was probably caused by the high cone of Rakata, which forced material from the collapsing eruption column northward. Reports of burns caused by hot ash from the area around Kalimbang in southern Sumatra provide evidence that some of the flows traveled to the northeast as far as 40 kilometers. It is clear that the injuries were due to horizontal flows rather than vertical ashfall because in one instance the survivors described hot gases and ash blowing up through the floorboards of a house.

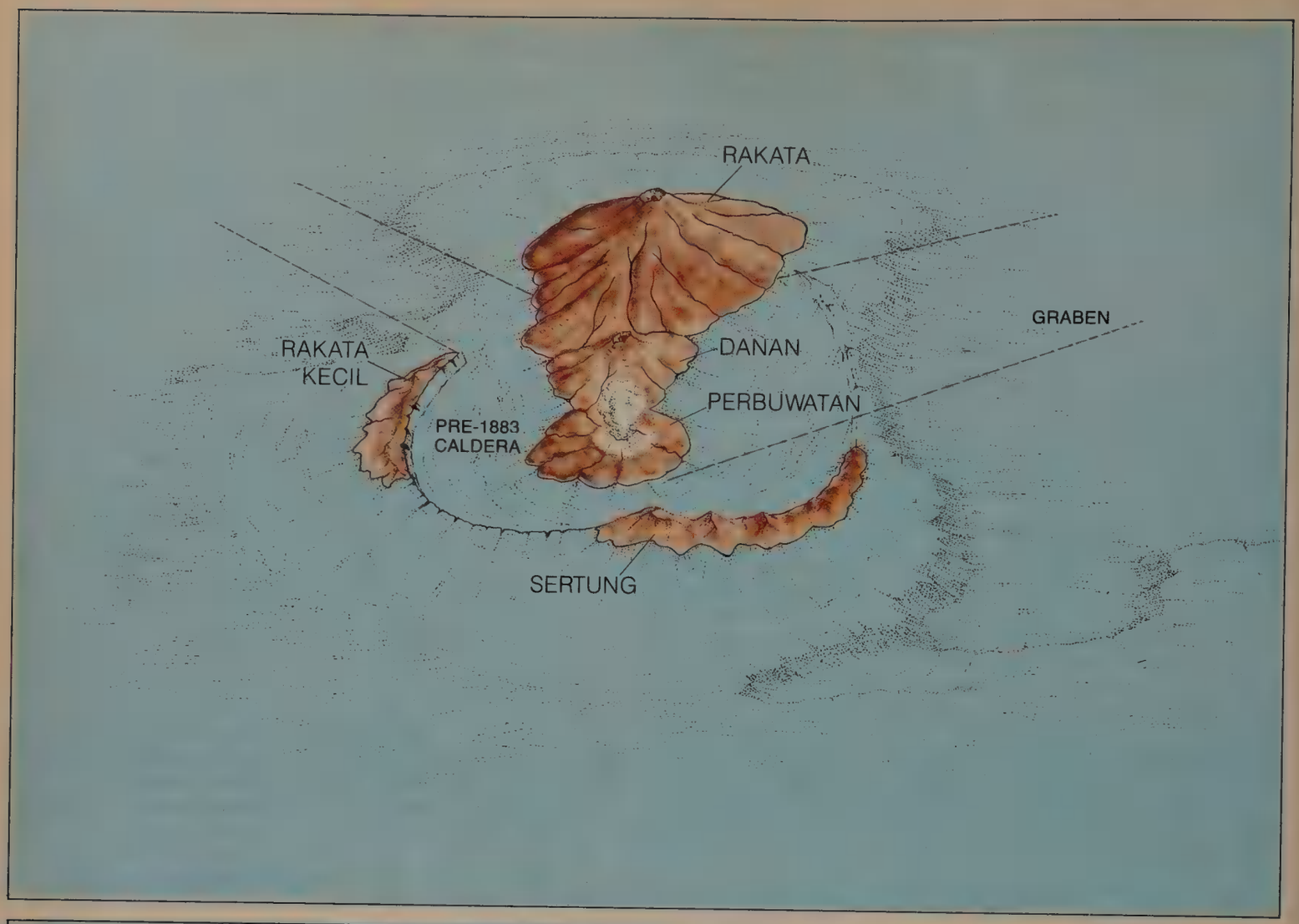
On August 28 the Dutch ship *Gouverneur-Generaal Loudon* attempted to sail from Telok Betong on Sumatra through the Sunda Straits north of Krakatau to Jakarta on Java. Parts of the straits had suddenly become too shallow to be navigable and were also blocked by floating islands of pumice. The ship was forced to deviate widely from its accustomed route, sailing west rather than east and finally passing south of Krakatau. The route the ship took roughly followed the outer edge of the ignimbrite that had been laid down on the floor of the sea.

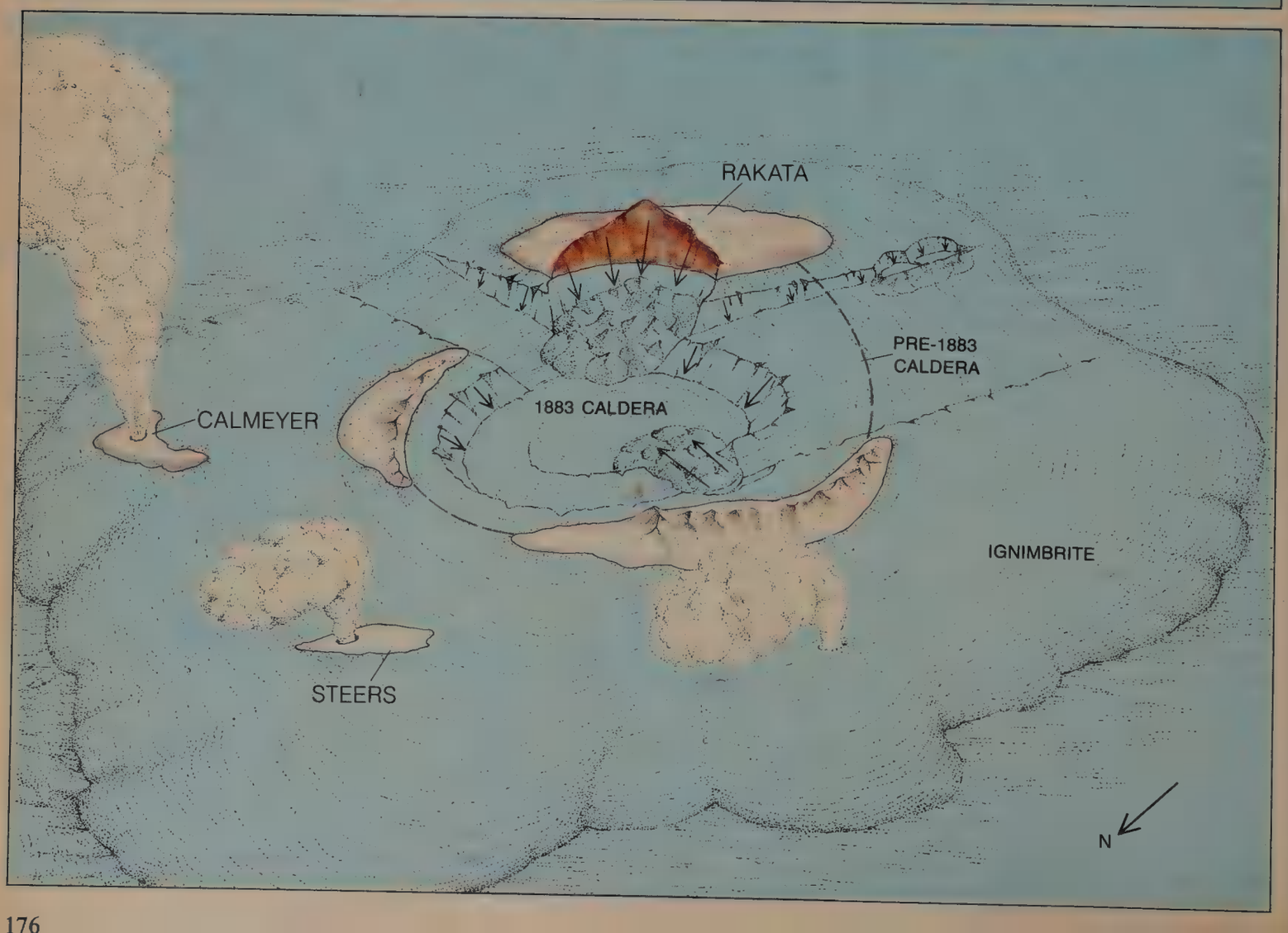
The detailed hydrographic charts Verbeek's team of investigators made in mid-October record the topographic changes caused by the eruption. The northern two-thirds of the island of Krakatau had disappeared. The coastlines of Sertung and Rakata Kecil had been extended by as much as three kilometers by deposits of pumice and ash. Even the southern remnant of Krakatau was girdled by deposits of white ignimbrite. Most of the volcanic material, however, had gone into the sea. Parts of the Sunda Straits originally from 20 to 60 meters deep had been filled by ignimbrite. In the straits to the north of Krakatau ignimbrite exposed above sea level had formed two islands, later named Steers and Calmeyer.

A new caldera, 290 meters deep at its southern end, had formed within the older and now partially filled prehistoric caldera. The distribution of the pyroclastic-flow deposits of the southern and eastern flanks of Rakata suggests they were laid down before the caldera formed, indicating that the magma chamber collapsed relatively late in the eruption sequence. Most of the island foundered into the sea when the roof of the chamber collapsed, but the cone of Rakata was apparently left perched on the southern rim of the caldera, its northern face virtually unsupported. This side of the volcano then slumped into the sea, leaving behind a spectacularly bisected volcanic cone, all that remains of the island of Krakatau.

One reason little progress was made until recently toward understanding what happened at Krakatau is that little attention was given to the physical characteristics of the volcanic deposits, particularly those left by the pyroclastic flows. Verbeek studied the petrology of the deposits and C. E. Stehn (who visited the area in 1927 when the eruption of a new, still-underwater volcano inside the 1883 caldera became perceptible at the surface) studied their stratigraphy. Howel Williams of the University of California at Berkeley made an important contribution in 1941, when he pointed out that much of the exposed pumice had been emplaced by pyroclastic flows rather than by the fallout of airborne material. Only within the past 15 years, however, have volcanologists begun to realize the significance of the physical characteristics of volcanic deposits, such as grain size and internal structure. These characteristics of the Krakatau deposits were studied for the first time when one of us (Self) and Rampino visited the island in 1979.

In many respects it is hardly surprising that these aspects of the deposits had


RAKATA
GRABEN
RAKATA
KECIL
DANAN
PERBUWATAN
PRE-1883
CALDERA
SERTUNG


RAKATA
PRE-1883
CALDERA
1883 CALDERA
CALMEYER
IGNIMBRITE
STEERS
N

not been examined before. Not only was most of the ignimbrite deposited in the sea but also the islands of ignimbrite, Steers and Calmeyer, were rapidly eroded and soon disappeared below the sea surface. The deposits that fell on the older islands were also rapidly eroded. Verbeek noted two months after the eruption that steep-sided gullies 40 meters deep had been cut into the deposits. The highly irregular topography was then enveloped in luxuriant and soon impenetrable vegetation. Today the deposits can be sampled only from boats at the bottom of crumbling sea cliffs. Yet it is on the physical characteristics of the deposits that many of the questions about the eruption turn.

Studies of the deposits provide the basis for our answer to the first question: What triggered the eruption? Major volcanic eruptions are caused by the sudden decompression of magma that is saturated or supersaturated with dissolved gases such as carbon dioxide and water vapor. The dissolved volatiles can be released in two ways. The pressure in the magma chamber may gradually increase as more volatiles come out of solution; when the pressure exceeds the strength of the overlying rock, the magma forces an opening to the surface. Alternatively, a tectonic process such as a landslide caused by an earthquake or movement along faults overlying the magma chamber may create an opening that causes instantaneous decompression. There is no evidence that any tectonic process was responsible for the eruption of Krakatau, and so an explanation must be sought in processes within the magma chamber itself.

Verbeek suggested that the final explosions began when seawater penetrated the magma chamber and reacted violently with the hot magma. This hypothesis was at one time widely accepted and was generalized to include the earlier explosions. Volcanic eruptions caused by the violent interaction of water and magma, called phreatomagmatic eruptions, do indeed take place either when magma heats and cracks rock close to an underground water table or when seawater somehow gains entry to the magma chamber. The mixture of groundwater or seawater and magma is highly explosive and usually gives rise to a distinctive deposit of very fine-grained and widely dispersed ash. The eruptions at Krakatau left rather different deposits that do not yield unequivocal evidence for this type of eruption.

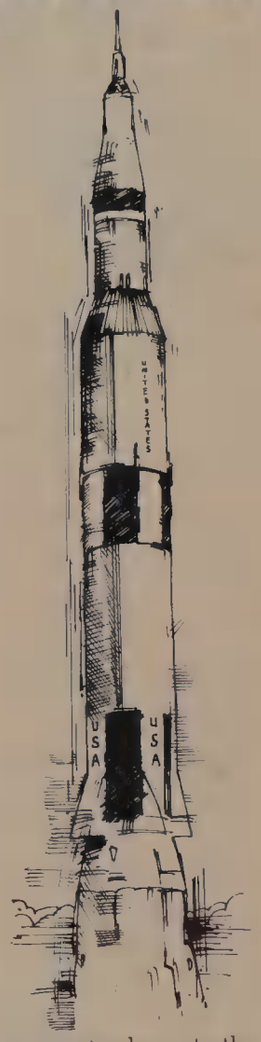
The intrusion of large amounts of seawater could be expected to cool the magma, so that the ejecta, particularly the finer-grained ash, would cool faster in a phreatomagmatic eruption than it would in other types of eruptions. The field evidence at Krakatau is somewhat equivocal on this point. On the one hand, the early air-fall deposits on Rakata Kecil, only 2.3 kilometers from the postulated vent, were so hot that the glassy pumice fragments were soft enough to weld together in places. The nature of this deposit thus makes it unlikely that the ejecta had been cooled by contact with water.

On the other hand, George P. L. Walker of the University of Hawaii at Manoa has argued that the character of some of the ignimbrite deposits suggests the magma had been cooled by water. The centers of the larger pumice clasts are frothier than the peripheries. In addition the clasts have an exterior skin such as might be formed by sudden cooling. These characteristics suggest the clasts were originally hotter than the surrounding matrix of ash and cooled from the outside in. It is, however, possible that seawater either cooled the erupted material as it left the vent or chilled the ignimbrite as it was deposited. (Most of the deposits visible today are at sea level.)

The geologic evidence therefore does not indicate clearly that there was direct explosive contact between seawater and hot magma in the chamber or conduit of Krakatau. Since the vent was close to sea level, it nonetheless seems probable that there were minor phreatic explosions, which may not have involved direct contact between the water and the magma, late in the sequence, beginning perhaps on the evening of August 26. These explosions may have weakened the roof of the main magma chamber, causing sudden decompression and the release of large volumes of vesiculating magma. We propose, however, that the

DRAWINGS OF KRAKATAU before and after the eruption are reconstructed from contemporary hydrographic charts. The island consisted of three volcanic cones: Rakata, Danan and Perbuwatan. The outlying islands Sertung and Rakata Kecil, and perhaps the southern edge of Krakatau itself, were remnants of the rim of a prehistoric caldera. Two submarine grabens (depressions) near the island indicate that the crust in the area was under extensional stress; this may have thinned the crust enough to accommodate the 1883 magma chamber. Most of the magma was ejected in the form of pyroclastic flows, which deposited on the sea floor a layer of ignimbrite (pumice fragments) as thick as 40 meters. The flows spread preferentially to the north and northeast, probably because the high cone of Rakata to the south acted as a barrier. The two new islands Steers and Calmeyer were areas of the ignimbrite exposed above sea level. Once the eruption had exhausted the supply of magma the chamber collapsed to form a new caldera and most of the island foundered into the sea. The 1883 caldera was probably elongate because the collapse followed the fault lines of the grabens. The drawings are not to scale.

Technology is in a state.



Tennessee is closer to the moon than it looks: much of the technology launched into space was researched in Tennessee. But space is just one side of Tennessee's technology story. Electronics and computer science are major industries. The University of Tennessee is becoming a nationally recognized energy research center, and we're making important breakthroughs in medical technology. When it comes to high technology, Tennessee is getting down to brass tacks and giving you the moon, and more. For more information, write or call Mike DuBois, Tennessee Department of Economic and Community Development, Andrew Jackson Bldg., Box 727, Nashville, TN 37219. 1-800-251-8594.

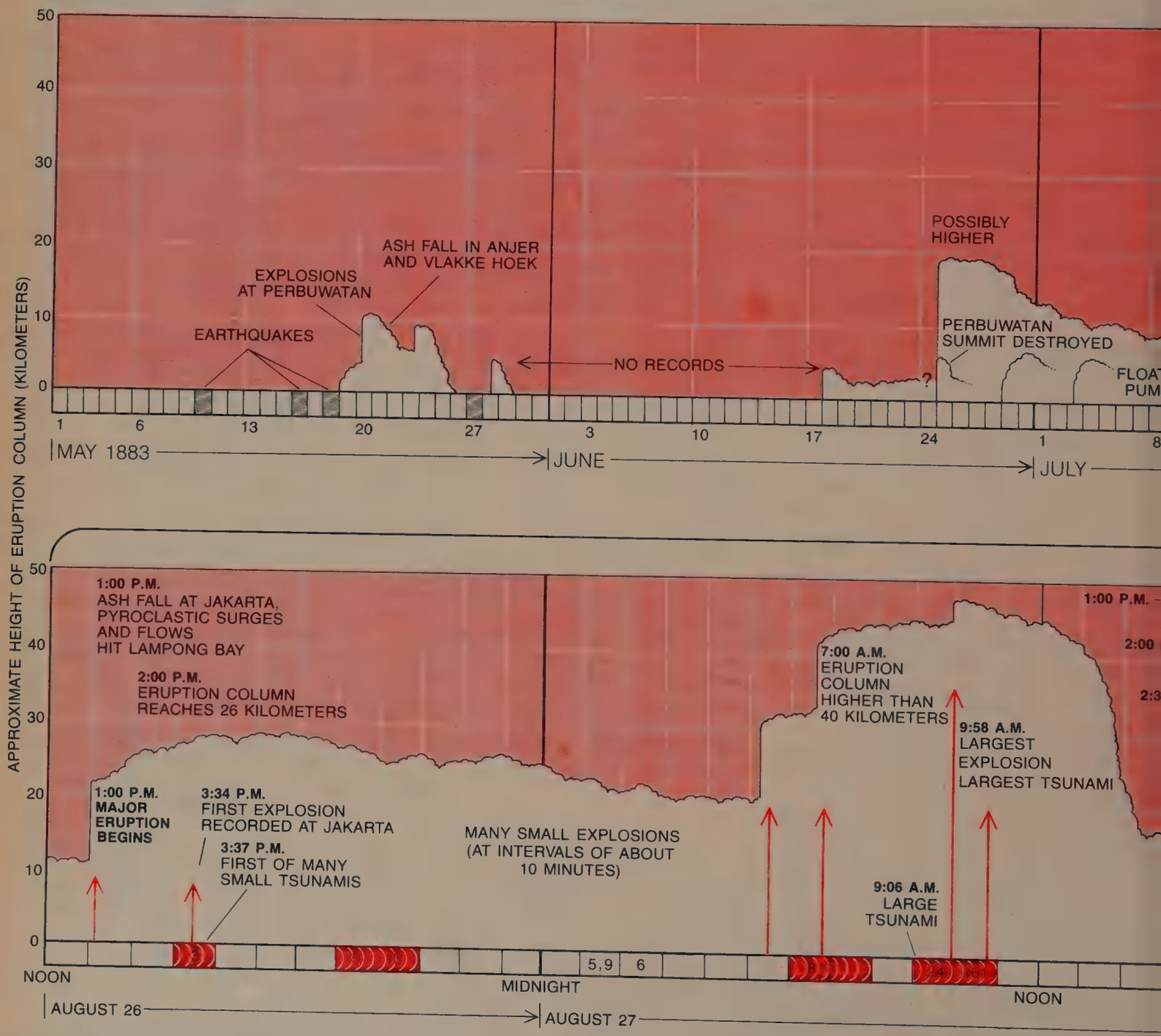
Tennessee
We get down to Brass Tacks.

interaction of water and magma was a contributory cause rather than the primary cause of the eruption.

The nature of the pumice fragments themselves points to a more likely triggering mechanism for the eruption. The frothy glass of the pumice includes a small number of mineral crystals that started to form in the magma chamber before the eruption. (When the magma cools relatively fast, glass is formed; slow cooling favors the formation of crystals.) The composition of a pumice and of any crystalline inclusions depends on the composition of the magma from which they formed. Magma ranges from basalt, a dark material relatively poor in silica, through andesite and dacite to rhyolite, a light gray material rich in silica. The lighter-colored magmas are called silicic; they form minerals with a high content of silica, such as quartz and feldspar. The darker magmas are called mafic; they form minerals rich in magnesium and iron, such as pyroxene.

Most of the Krakatau deposits consist of pale dacitic pumice that contains relatively few mineral crystals. Some of the deposits are strikingly different, however; streaks or bands of darker glass are intermixed with the lighter glass, and some clasts are entirely dark. The darker pumice also includes crystals with compositions different from those in the lighter pumice. Such mixed pumices are by no means rare in pyroclastic deposits. It has been suggested that they may be formed when hot basaltic magma intrudes into light rhyolitic or dacitic magma and the two mix.

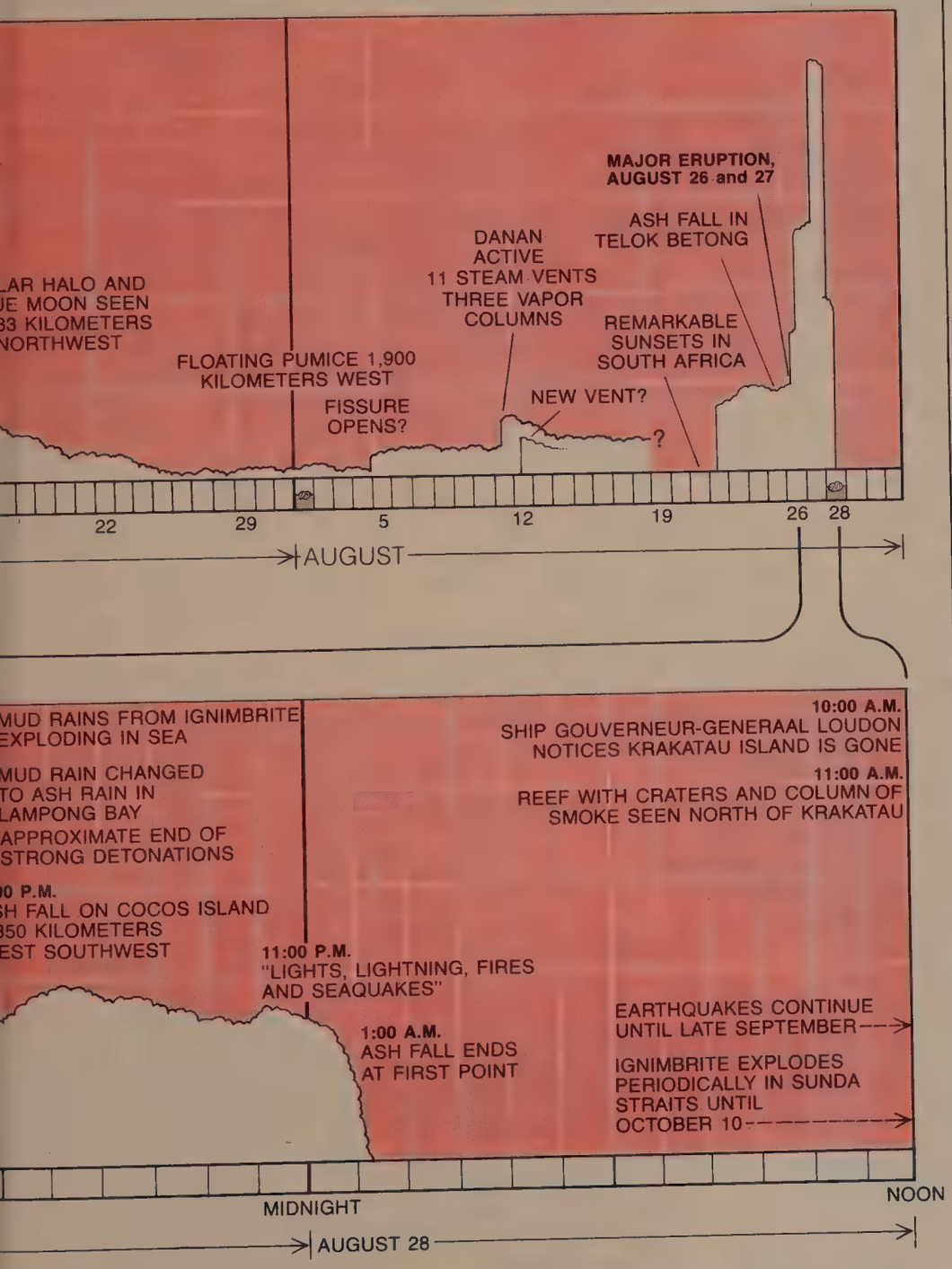
The intrusion of fresh hot basaltic magma into the base of a chamber of dacitic magma may cause a violent convective overturn within the chamber. Silicic magmas tend to be less dense than more mafic magmas, largely because of the differences in their composition. Most dacitic magma chambers are thought to be stratified according to composition, with the most silicic magma at the top and the somewhat denser, more mafic material lower down. Such a system is stable. Stephen Sparks and Herbert Huppert of the University of Cambridge have suggested, however, that a pulse of hot basaltic magma that intrudes into the chamber may initiate convective overturn by superheating the



ERUPTION SEQUENCE OF KRAKATAU has been pieced together from many sources: diaries and notebooks kept by Dutch administrators and others living on the neighboring islands, logs kept by the officers on watch on ships in the Sunda Straits, events recorded by a pressure gauge at a gasworks in Jakarta and by tide gauges along the Sunda Straits and occasional observations made by people

adjacent layer, causing it to become less dense than the overlying silicic magma. A second possibility is that the hot basaltic magma initially rests quietly on the floor of the magma chamber and violent overturn takes place only when it cools. As the basaltic magma cools, volatiles come out of solution and crystals settle out, decreasing the density of the remaining liquid magma until it is lower than that of the overlying layers.

Why should convective overturn initiate an explosion? In general the eruption of basaltic magmas tends to be less explosive than that of more silicic magmas because they are less viscous and the bubbles of gas created by volatiles coming out of solution easily percolate through or effuse out of them, whereas the bubbles formed in more viscous silicic magmas are trapped there and cause explosive fragmentation of the magma when their internal pressure exceeds its strength. In an impressive series of laboratory experiments, however, J. Stewart Turner of the Research School of Earth Sciences of the Australian National University and Sparks and Huppert have demonstrated the importance of the volatile content of the newly introduced magma to the process of magma-chamber mixing. Fresh basaltic magma may be rich in volatiles kept in solution by the pressure at the depth at which the magma formed. Convective overturn carries the volatile-rich material to shallower levels in the chamber, where the ambient pressure is lower. As the magma rises, the volatiles rapidly come out of solution, the pressure with-

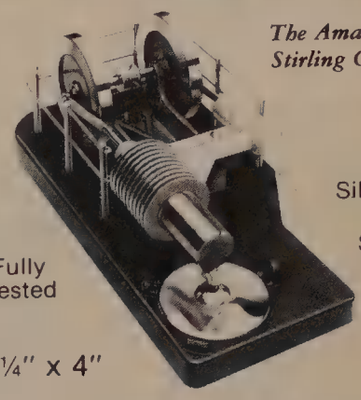


as far away as South Africa. Numbers in colored boxes give the relative magnitude of the tsunamis. The relative magnitude of the explosions is indicated by the height of the arrows. Only a few of the many explosions that occurred in the course of the eruption sequence are shown.

Hot Air Engine

Heat is the only fuel required!

The Amazing Stirling Cycle



Silent

Safe

Fully Tested

7¼" x 4"

Speeds in excess of 1000 RPM can be attained from this miniature Engine. This ALL-METAL, precision machined, Engine comes complete with built-in alcohol burner, red flywheels and chassis mounted on a green base.

Each Engine comes completely assembled and ready to run, accompanied by our lavishly illustrated, 128 page book entitled "Stirling Cycle Engines".

Engine and book $53.50

Send check, money order or use Visa, American Express or MasterCard.
Credit Card Orders: (800) 528-6048

Solar Engines
Dept. SA
4020 E. McDowell Rd., Phoenix, AZ 85008

Discover the world's oldest hobby.



It really isn't hard to build a fine quality, wooden ship model. The secret's in our kits—imported from Europe—where they are hand crafted in walnut and rich woods by the finest model makers in the world. Finished brass, bronze and wood fittings are used throughout. The hard work is done for you—most wooden parts are precut. Assembly is simple as you follow step-by-step instructions and large scale plans. Everything you need to complete your model is included, and we guarantee satisfaction.

Discover ship modeling, now! Send for our full-color, 72-page catalog illustrating over 50 model kits priced from $25 to $700. Only $3, refunded with your first order.

Mail this coupon today!



Dept. SA-113

Model Expo, Inc.
23 Just Road,
Fairfield, N.J. 07007

Please send full-color catalog; $3.00 enclosed.

Name _____

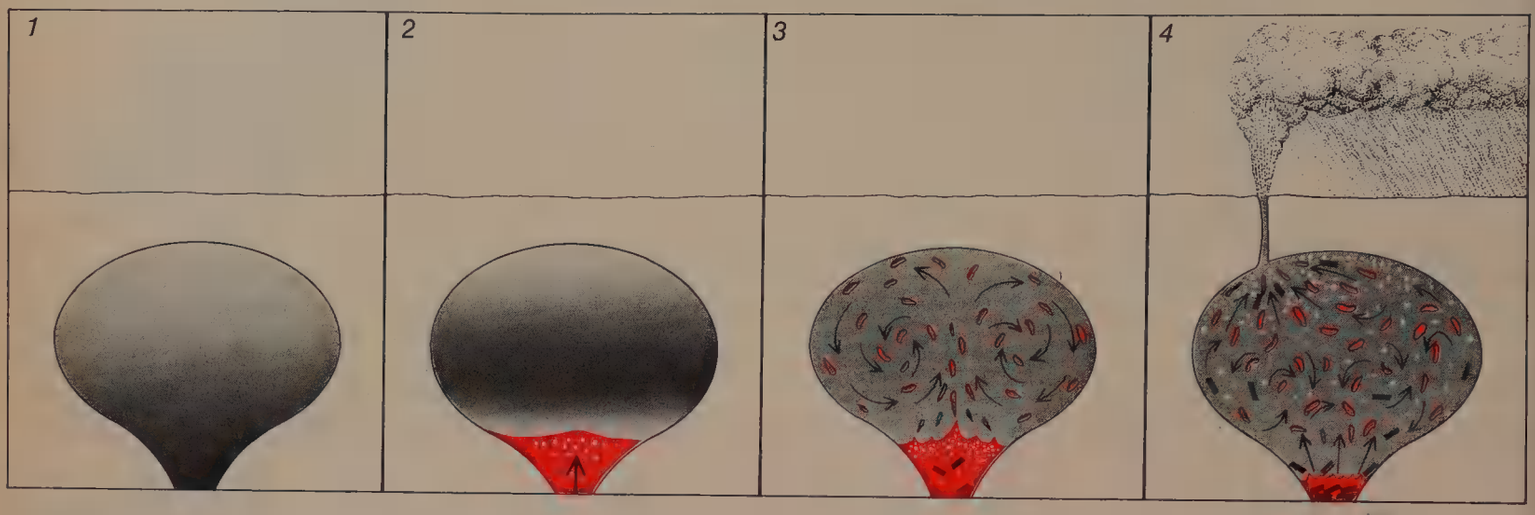
Address _____

City _____

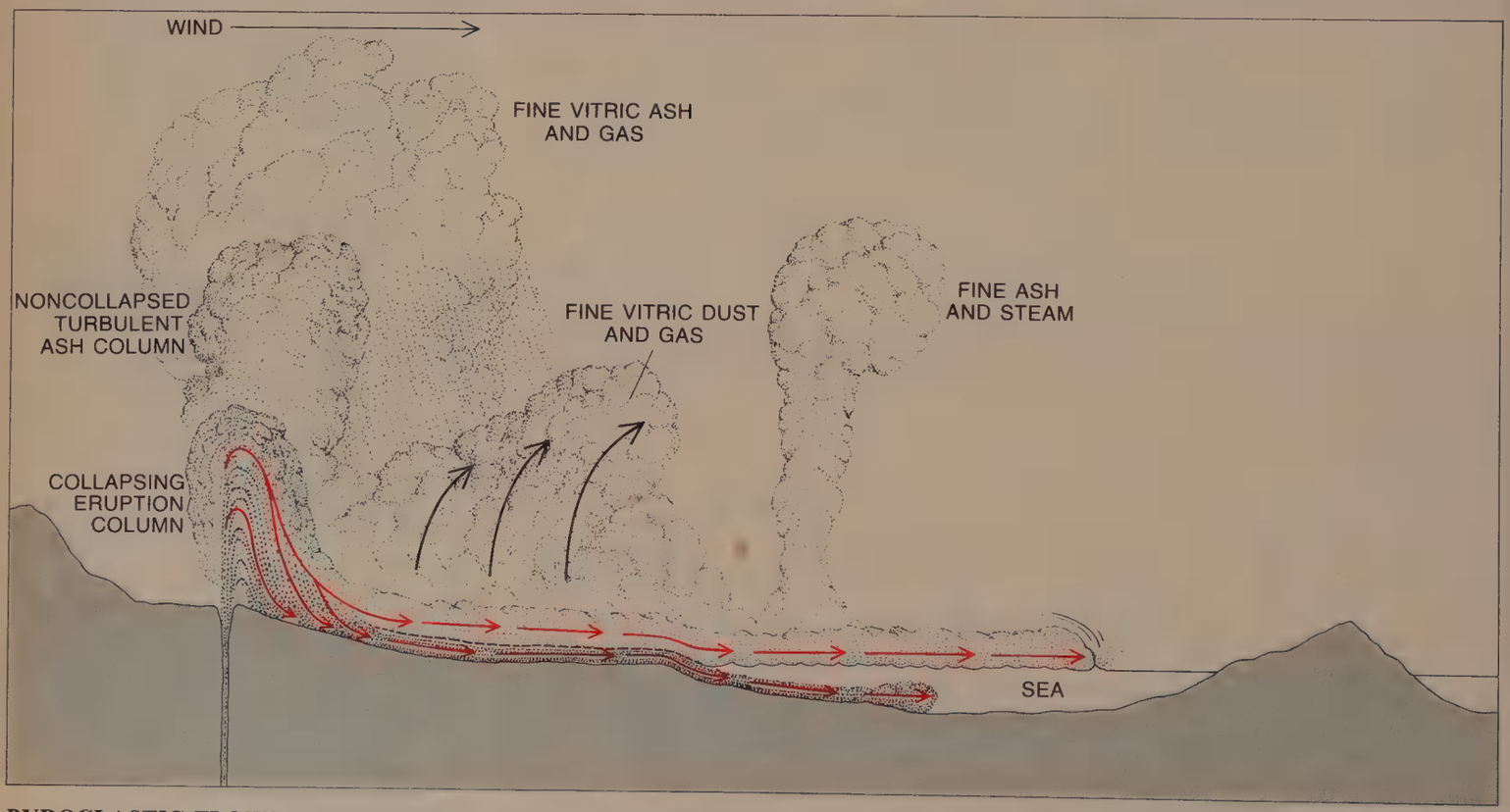
State _____ Zip _____

in the magma chamber increases and an explosive eruption is likely to follow.

An important piece of historical evidence supporting the hypothesis that the eruption of Krakatau was caused by magma mixing came to light during the preparation of this article. On May 27, 1883, soon after the first explosions of Perbuwatan in May, a party of no fewer than 86, including the Dutch mining engineer J. Schuurman, visited Krakatau. Schuurman wrote a detailed account of his visit, noting that 60 centimeters of dark "ash" overlay 30 centimeters of lighter-colored "pumice." He also collected samples of the ejecta, which were later analyzed. The samples of pumice are similar in composition to the dacitic pumice ejected in the August explosions. The gray ash, however, is basaltic in composition. It seems likely, therefore,



VIOLENT MIXING OF MAGMAS of differing composition may have triggered the eruption of Krakatau. Magma ranges in composition from silica-poor basalt through andesite and dacite to silica-rich rhyolite. The less silica there is in the magma, the denser it tends to be. The mixture of dacitic magmas in the chamber of Krakatau may originally have been stratified according to density and therefore been stable (*1*). A pulse of fresh basaltic magma, hotter than the material already in the chamber and also rich in volatile substances, may have intruded into the bottom of the chamber, and at first it may have rested quietly on the chamber floor (*2*). The hot magma would superheat the layer of dacitic magma immediately above it, however, making that layer less dense than the overlying layers. At the same time, as the basaltic magma began to cool, crystals would settle out and volatiles would begin to come out of solution, making the basaltic magma itself less dense than the lower layers of dacitic magma. Either or both of these processes would cause violent convective overturning of the material in the chamber (*3*). As the volatile-rich basaltic magma rose to shallower levels of reduced ambient pressure, dissolved volatiles would rapidly come out of solution. The pressure within the chamber might therefore increase to the point where the magma would force its way to the surface and erupt (*4*). Mixed pumices found at Krakatau suggest this mechanism caused the eruption.



PYROCLASTIC FLOWS are masses of incandescent volcanic material fluidized (given a density and a viscosity lower than those of an aggregation of solid material) by hot gases. As the flow travels it segregates into a dense lower zone in which the flow is generally laminar and a light upper zone in which the flow is turbulent. When the flow reaches the sea, the denser material may plunge into the water, but the upper part of the flow may temporarily have a density less than that of seawater and therefore may travel on the surface. As the fluidization dissipates, the remaining material becomes denser and more of it enters the sea. The process continues until only a turbulent cloud of ash and steam is left. The hot material can cause secondary explosions as it sinks into the sea. It has been argued that the grading process that takes place within the flow itself and the winnowing of the ash by the secondary explosions produce a fine, crystal-poor ash that is distributed over a wide area by the secondary explosions. Rogier D. M. Verbeek reported finding deposits of this type near Krakatau.

sands of square kilometers. The distinctive feature of these ash deposits is that their thickness and their grain size vary little with distance from the source, a distribution that suggests the ash was generated by explosions of remarkable power.

The Rotoehu deposits erupted from the Okataina volcanic center in New Zealand some 50,000 years ago yield evidence that such ash deposits may also be crystal-poor. Walker argues that the crystals become separated from their parent pumice fragments by a natural grading process. As the flow moves toward the sea the denser crystals sink and the pumice dust is lofted. Further winnowing of the crystals from the dust accompanies the explosions that occur when the flow enters the sea, as powerful turbulent vortexes sifting through the cloud of ash created by the explosions carry away the lighter dust.

At this remove in time it is, of course, difficult to determine exactly what happened when the Krakatau pyroclastic flows entered the sea. Contemporary illustrations, however, clearly show large circular craters on Calmeyer, one of the two islands created by the deposition of ignimbrites. The craters strikingly resemble those made by the secondary explosions that occurred when pyroclastic flows erupted by Mount St. Helens in May, 1980, entered Spirit Lake. Moreover, Verbeek noted that the uppermost layers of ash on the island of Sebesi were extremely fine-grained and consisted almost entirely of glass. This fine crystal-poor ash may well have been the fallout from powerful secondary explosions. For these reasons we suggest that although the gigantic blasts at Krakatau up to 9:58 A.M. on August 27 were probably caused by the expulsion of the pulses of magma that generated the pyroclastic flows, at least some of the many powerful explosions in the eruption sequence were explosions caused by pyroclastic flows entering the sea.

It was the great tsunamis set in motion by the eruption that wreaked the most havoc. Even though the tsunamis have been studied in detail by many investigators, their cause or causes remain a matter of debate. Tsunamis are generally caused by the sudden vertical movement of the sea floor, usually as the result of an earthquake. As late as the 1960's it was argued that the Krakatau tsunamis were caused by the similar mechanism of caldera formation. Discussion of this point has centered on the question of when and how quickly the roof of the magma chamber collapsed and on the question of whether the tsunamis were preceded by a lowering of the sea level, such as would be expected when water rushed in to fill the newly formed caldera. The question of the exact timing of the caldera collapse has not been settled, but it seems to have

How to get 336 lenses in your camera bag.



You have two choices. Either you find a bag that's extraordinarily large. Or you find one lens that delivers 336 different focal lengths. Like the 75-410mm zoom lens system from Vivitar.

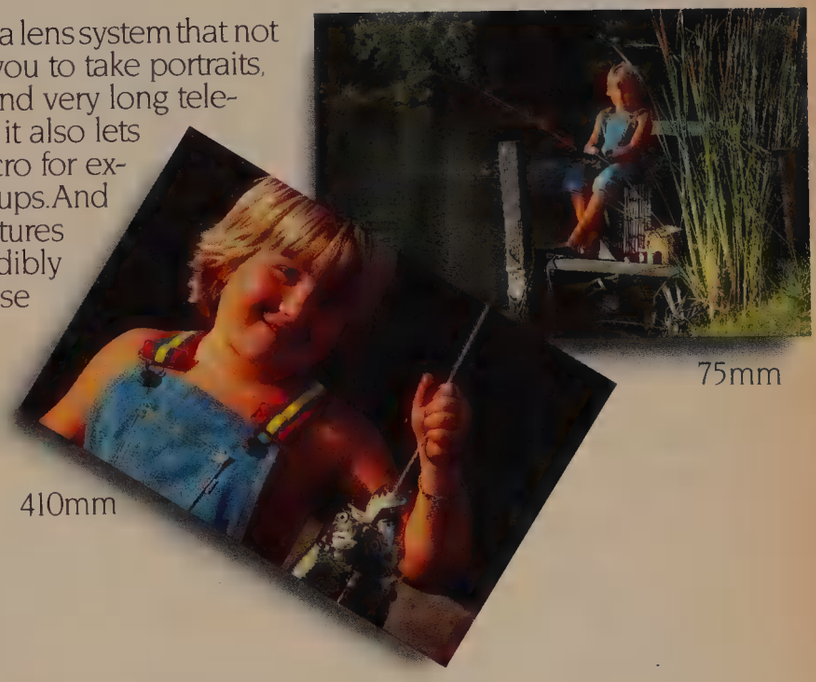
Vivitar

Our lens system is made up of two components. We start with a 75-205mm f/3.5-4.5 lens, which is the fastest, smallest and lightest lens in its class. And to it, we add a specially designed 2X Matched Multiplier™ for doubling the focal range.

The result: a lens system that not only allows you to take portraits, telephotos and very long telephoto shots, it also lets you use macro for extreme close-ups. And all of the pictures will be incredibly sharp because of the lens' high degree of contrast and resolution.

But performance is not the only part of our system that's unsurpassed. So is its value. There's simply no better 75-410mm lens system at any price. And each system is backed by a 5-year limited warranty.

So the choice is quite simple. You can buy 336 lenses for your Canon, Minolta, Nikon, Olympus, Pentax or other popular 35mm SLR camera. Or you can buy one Vivitar 75-410mm zoom lens system. Ours is certainly a nicer fit in your camera bag.



The 75-410mm Zoom System by Vivitar

For a brochure (U.S. only) on our complete family of photographic products, send a self-addressed envelope with 40¢ postage to: Vivitar, 206AD, Box 2100, Santa Monica, CA 90406. Vivitar Corporation 1983. In Canada: Vivitar Canada Ltd./Ltée.

What if you chose as a technical



"At Union Carbide, an HP computer network automates production management and has paid for itself in less than two years."

Union Carbide's Linde Division in Tonawanda, N.Y., manufactures custom production equipment. Factory systems manager Doyce Coffman says, "When production demand fluctuates, it complicates our scheduling, product costing, and long range planning.

"To solve these problems we automated production and materials management with a network of HP 1000 and 3000 computers and 43 data capture terminals.

"Because these systems continuously monitor operations and respond in real time, we can identify and eliminate production bottlenecks on line. The payback? Improved efficiency factory-wide and a dollar savings that covered the cost of our entire HP network in less than two years."



What should you expect from a working partnership with Hewlett-Packard?

Results with assurance! DataPro Research Corporation's annual survey of computer users has ranked Hewlett-Packard *number one* in Customer Support Satisfaction for three years in a row. Why? Because the same quality and reliability engineered into HP products is built into our customer support programs.

From the outset, a Local Support Team of specialists who understand your needs will help you develop a cost-effective strategy to meet your goals—from selecting the right equipment to training your people.

You can tie your HP system into one of our Technical Centers where HP specialists will monitor your system's performance. Our routine, remote diagnostic tests can help iden-



The HP circle of quality customer support

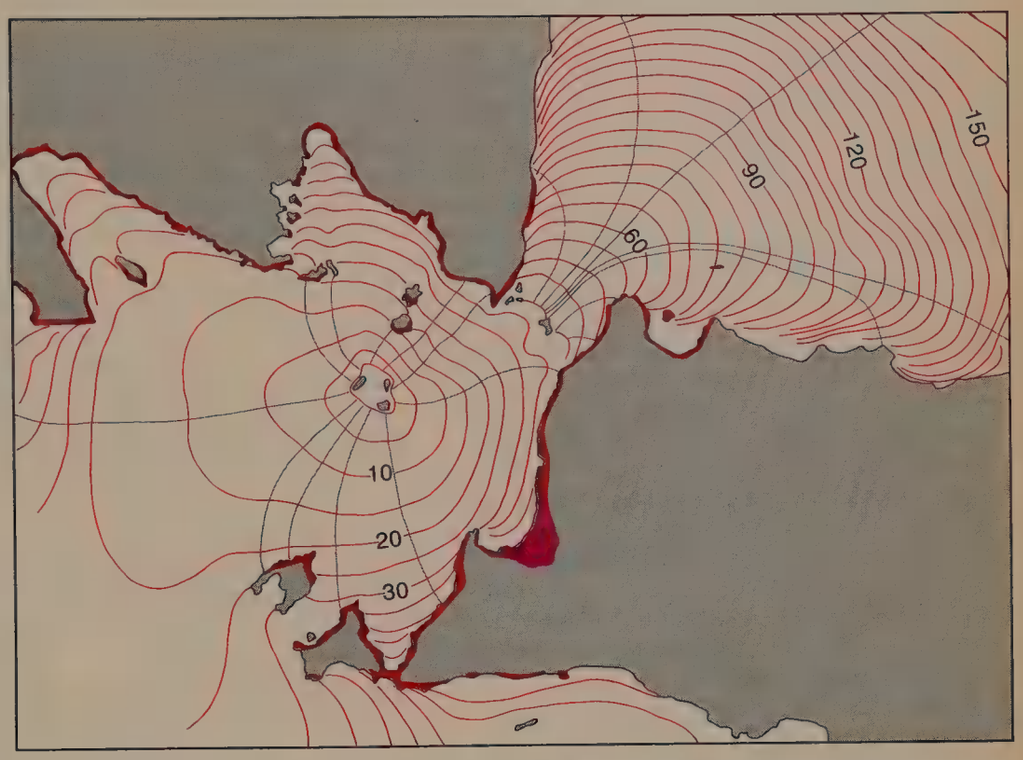
come late in the sequence, perhaps even after many major tsunamis. The evidence on the second question is also somewhat equivocal. At many points along the Sunda Straits the arrival of most of the tsunamis was marked by a rise in the sea level. A few retreats were also apparently recorded, however, and most of the tide gauges were so far away from Krakatau that an initial retreat may have been too small to register on them.

There are three alternative causes for the tsunamis. Some workers have suggested that submarine eruptions, which like a depth charge would create an outward-propagating water dome, may have caused some of the tsunamis. Verbeek suggested in 1884 that the largest tsunami was set in motion by the slumping of the northern half of the volcanic cone Rakata into the newly formed caldera. Verbeek also suggested that the sudden displacement of water by volcanic ejecta "falling" into the sea could have set the waves in motion. Given that many cubic kilometers of material entered the sea in the form of pyroclastic flows, this possibility seems to us the likeliest one.

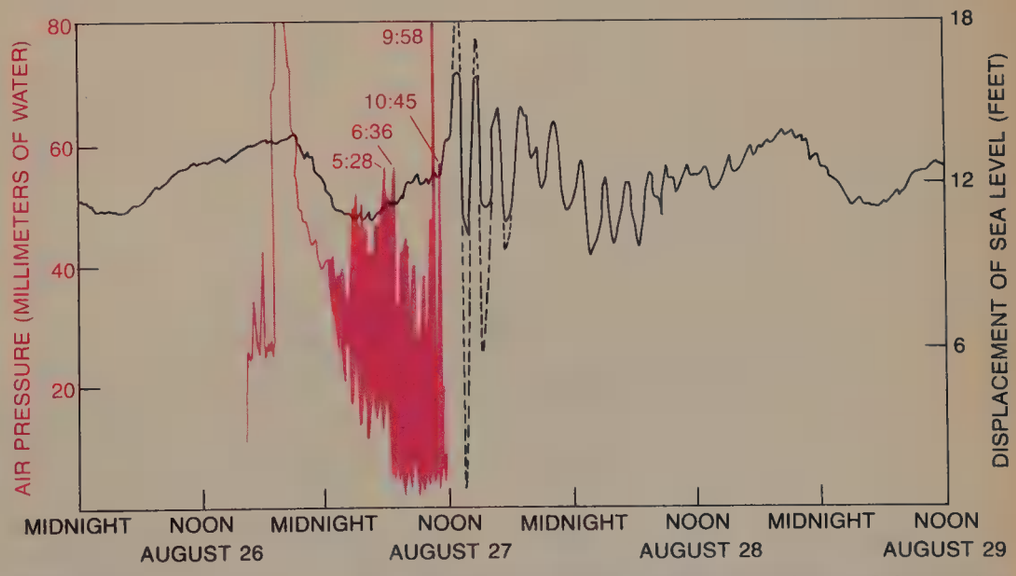
The correlation of the time of arrival of the waves at points along the Sunda Straits with events in the eruption sequence is critical to arguments about the origin of the sea waves. Establishing the chronology has been complicated by two problems. First, not all the tsunamis were true tsunamis, that is, waves that had traveled entirely by water. Second, it has proved difficult to synchronize the eruption sequence with the timing of the atmospheric disturbances recorded by the gasworks pressure gauge at Jakarta and sea waves recorded by the tide gauges along the straits.

Many of the sea waves, particularly those recorded at points far from the volcano, appear to have traveled extremely fast: at the speed of air waves rather than at speeds typical of sea waves. The velocity at which a wave moves through water is proportional to the depth of the water. More precisely, the velocity of the wave is equal to the square root of the acceleration of gravity multiplied by the depth of the water. The waves from Krakatau reached Honolulu in only 11 hours, a travel time that implies an average water depth of 17 kilometers. The average depth of the deep ocean is much less: about four kilometers. Furthermore, sea waves attributed to the eruption were observed in parts of the oceans where they could not reasonably be expected to appear; for example, waves were detected on the far side of island chains, a barrier through which they could not possibly have propagated.

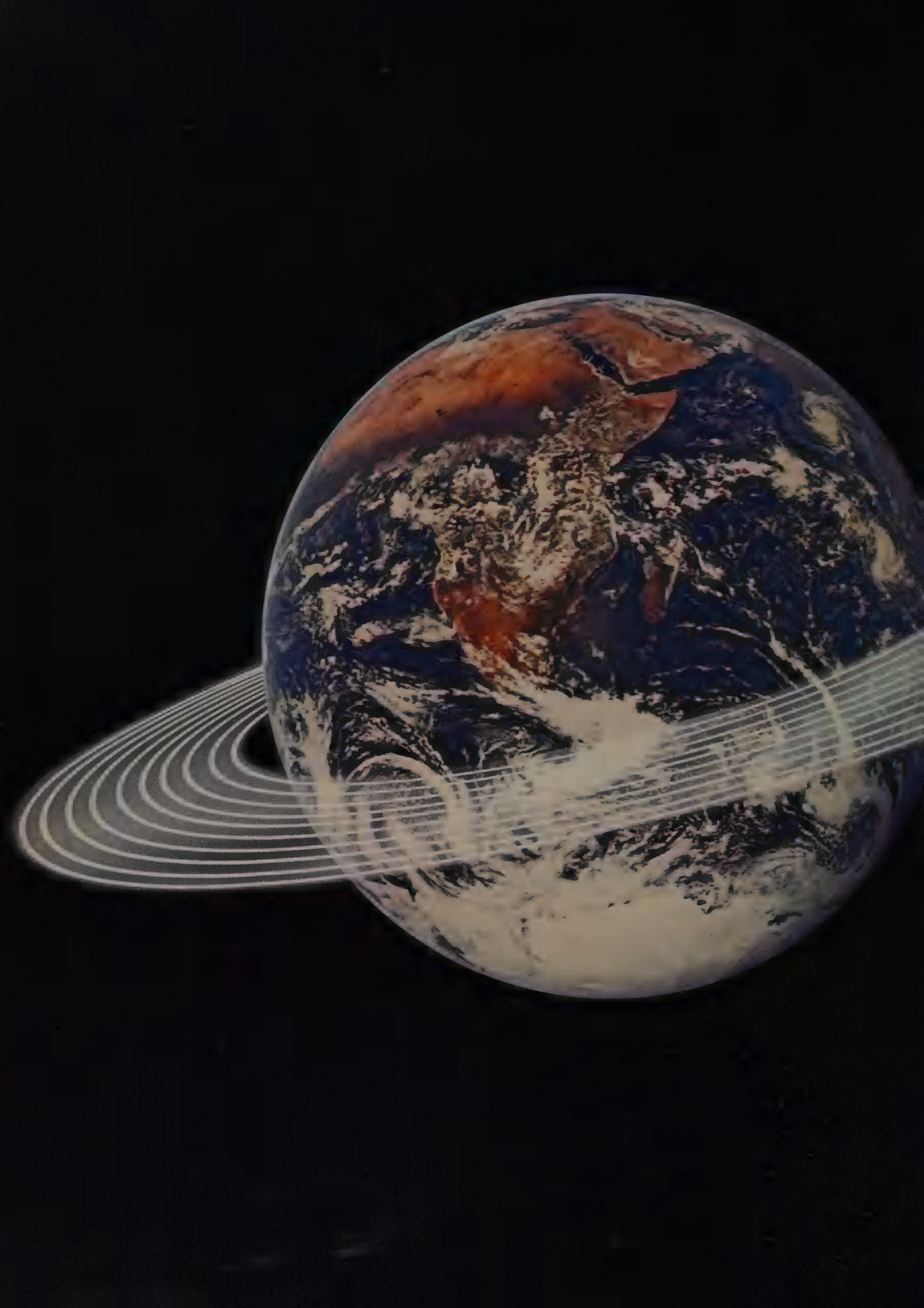
These discrepancies came to light soon after the eruption but for many years were unexplained. In 1955, after microbarographs capable of registering small atmospheric pressure variations were developed, Maurice Ewing of the Lamont-Doherty Geological Observatory and Frank Press of the Massachusetts Institute of Technology were able to demonstrate that seismic waves through the solid earth can be coupled to the atmosphere. In 1967 David G. Harkrider of Brown University and Press showed that the long trains of pressure pulses set up in the atmosphere by nuclear explosions transfer some of

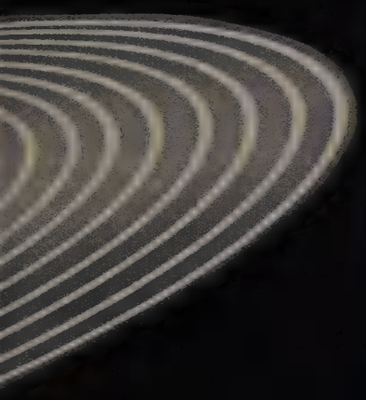


MAP OF TSUNAMI TRAVEL TIME is based on the times of arrival of the tsunamis recorded by tide gauges at various points along the Sunda Straits. The arrival times of the tsunamis were distinguished from those of the trains of secondary waves set up when the tsunamis entered the bays where most of the tide gauges were installed. Then the tsunamis were traced backward in time; their speed at each point in time was determined from the depth of the water through which they were then moving. Contour lines (*colored lines*) were derived from the paths and travel times of a number of tsunamis (*gray lines*), making it possible to find the travel time of a tsunami from Krakatau to any point on the coast. The inundated areas are in color.



PRESSURE AND TIDE GAUGES at Jakarta recorded the major explosions and the ensuing tsunamis on August 27, 1883. A pressure gauge at the gasworks (*color*) fortuitously recorded the arrival of the atmopsheric pressure waves generated by the explosions. The broad peak in the trace on August 26 is unrelated to the eruption; it may have been caused by an increase in the pressure of the gas in the tank the gauge was monitoring. The tide gauge at Tanjong Priok (Jakarta's harbor) (*black*) recorded the arrival of the sea waves set in motion by the same explosions. Some of the waves were so high that they exceeded the range of the gauge. The broken-line peaks are estimates of the displacement in sea level based on various observations, such as the number of steps in a flight of stairs leading down to the harbor that were covered by water. The highest peak corresponds to the tsunami that arrived at Jakarta at 12:16 P.M. The peaks following it include secondary waves set up in the harbor and the straits by the tsunami.





Who on earth will install enough lightguide fiber to circle the earth 10 times?

Western Electric, an AT&T company.

In just five years, Western Electric has already created and installed enough lightguide fiber to circle the earth. That's 25,000 miles of fiber in 60 lightwave communications routes.

Those totals should increase at least tenfold by the year 2000, as Western Electric continues to install lightwave systems to transmit the high-speed voice and data services of the Information Age.

Far more efficient and economical than copper cable, lightwave communications speeds your messages over 90 million light pulses per second through hair-thin glass fibers. A single lightguide cable can carry 240,000 telephone calls at once.

Right now Western Electric is installing the world's largest lightwave system in the busy Northeast Corridor. Another will transmit the entire 1984 Los Angeles Summer Olympics to ABC television studios.

And our research and development arm, Bell Labs, is working on a transatlantic lightwave cable that will triple calling capacity and enhance transmission quality.

These systems are forerunners of a continually expanding lightwave network that one day will literally circle the earth, bringing everyone exciting new information services.

Western Electric, an AT&T company. Applying the technologies of lightwave, microelectronics, and software to make the dream of the Information Age a reality.

AT&T

their energy to the ocean, causing sea waves in regions far from the explosion.

The coupling mechanism is complex; the amount of energy transferred depends strongly on resonance, that is, on the natural vibrational frequencies of the atmosphere and the ocean. Nevertheless, the general principle that governs the interaction can be stated simply: The ocean can be thought of as responding to alternating increases and decreases in atmospheric pressure in such a way as to maintain hydrostatic equilibrium. Increased atmospheric pressure causes the ocean to become shallower and decreased pressure causes it to become deeper, so that the weight of the ocean-atmosphere column above each unit area remains as nearly constant as possible. Harkrider and Press recognized the analogy between a nuclear explosion and a volcanic one and suggested that the same kind of phenomenon might explain the anomalous tsunamis that followed the eruption of Krakatau.

Much patient detective work was still needed to determine what caused the true tsunamis that inundated the coastlines along the Sunda Straits. The arrival of sea waves recorded by tide gauges at various points along the straits had to be distinguished from the arrival of secondary waves set up by the tsunamis within the enclosed bodies of water, such as bays, where most of the tide gauges were installed. Then the tsunamis had to be traced to their point of origin, taking into account refraction, or changes in velocity, caused by the varying depth of water in the straits. In 1981, I. Yokoyama of the Usu Volcano Observatory in Japan analyzed the refraction of the Krakatau waves and published a map of the travel time of the tsunamis. On the basis of the map he concluded that some of the minor tsunamis were probably caused by volcanic ejecta entering the sea but that the largest tsunami, which reached Jakarta at 12:16 P.M. on August 27, must have been caused by an underwater eruption.

John H. Latter of the New Zealand Department of Scientific and Industrial Research came to somewhat different conclusions in a paper published the same year. Latter set out to establish the chronology of events as accurately as possible from the records kept by both the pressure gauge at the Jakarta gasworks and the tide gauge in Jakarta's harbor. He calculated the times of explosions at Krakatau from the times of arrival of atmospheric pressure waves at Jakarta, taking into account the atmospheric travel time of eight minutes and the five-minute difference between Jakarta time and Krakatau time. He then showed on the basis of the timings of the explosions and the map of tsunami travel times worked out by Yokoyama, according to which the travel time for a tsunami traveling from Krakatau to Jakarta was two hours 25 minutes, that the chronometer on the tide gauge at Jakarta was inaccurate by three and a half minutes with respect to the pressure gauge. Having established the chronology of events, Latter was able to show that the arrival times of atmospheric waves and sea waves at Jakarta caused by some of the minor events of the eruption sequence were correlated.

In several instances, however, the sea wave could not be correlated with a matching atmospheric wave and thus with the explosive event that caused both. The largest atmospheric wave, caused by the massive explosion heard as far away as the Indian Ocean, reached Jakarta at 10:08 A.M. If the atmospheric travel time and other factors are taken into account, it must have been caused by an explosion taking place at Krakatau at 9:58 A.M. A large sea wave was recorded on the tide gauge at Jakarta at almost exactly the same time as the atmospheric pressure wave was recorded at the gasworks. Given that the travel time of a sea wave from Krakatau to Jakarta is approximately two hours 25



ANAK KRAKATAU ("Child of Krakatau") is a new volcanic cone that is building up in the caldera left by the 1883 eruption. It is at roughly the same site where Danan once stood. The new cone broke the surface of the sea late in January, 1928. Successive eruptions of Anak Krakatau have since largely filled the caldera left by the 1883 eruption with debris. The photograph was made by Maurice Krafft.

minutes and that Latter could find no evidence for an event taking place at Krakatau at about 7:40 A.M., he concluded that the large sea wave was caused by the coupling of the atmospheric wave to the water. The true tsunami caused by the explosion at 9:58 A.M., which had a much larger amplitude than the false tsunami, did not arrive at Jakarta until 12:16 P.M.

At this point the details of the timing become all-important. Given the sea-wave travel time, the large tsunami that arrived at 12:16 should have originated at Krakatau at about 9:45 A.M., that is, before the major explosion took place. Latter was confident of the accuracy of his meticulously checked timings, and he was certain there must be a causal link between the largest air and sea waves, and so he concluded that the tsunami originated not at Krakatau itself but rather at a point closer to Jakarta's harbor by a distance equivalent to the time discrepancy. He then used Yokoyama's map to show that the point of origin of the tsunami set in motion by the great explosion at 9:58 A.M. was 10 to 15 kilometers from Krakatau. This point corresponds roughly to the outermost edge of the newly formed island Calmeyer. Similar analyses of the timing of paired explosions and sea waves led Latter to conclude that at least three of the Krakatau tsunamis were probably caused by massive pyroclastic flows advancing into the sea, such as the flow that formed Calmeyer.

Although the Krakatau tsunamis are the largest tsunamis known to have been generated by pyroclastic flows, the same mechanism has been shown to be responsible for tsunamis following the eruption of other volcanoes. In 1980 Juergen Kienle and Samuel E. Swanson of the U.S. Geological Survey reported that the nine-meter tsunami that swept through English Bay in Alaska after the eruption of the volcano Augustine in 1976 was set in motion by a pyroclastic flow entering the bay.

Since the eruption of 1883 a new volcano, Anak Krakatau ("Child of Krakatau"), has been building up at a site roughly corresponding to the position of the old volcanic cone Danan. The new volcano was initially under water, but the cone emerged late in January, 1928. Since then successive eruptions have gradually filled the northern sector of the 1883 caldera with lava and pyroclastic flows. The focus of volcanic activity is shifting southward and may well be following the same fissure that determined the alignment of the volcanic cones on the original island. A large island, aligned like the old one northeast to southwest, may eventually develop. It seems that a new Krakatau is rising like the phoenix from the ashes of its own destruction.

Flatland

By Edwin A. Abbott



FLATLAND is one of the classics in the field of science fiction. It is a satire of great sensitivity, depicting a world of two-dimensional creatures unable to visualize a third dimension. They are intolerant of less advanced civilizations unable to comprehend their complex 2-dimensional society. Originally published in 1884, FLATLAND anticipates the concept of a 4th dimension. This book will appeal both to lovers of science fiction and to math buffs. A new introduction by Karen Feiden is included. ILLUSTRATED.

$9.95 *plus* **$1.25** handling

How to Argue and Win!



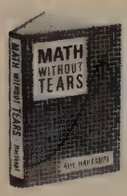
Here is a clear simply written basic guide to logical thinking, showing how to spot the fallacies, the prejudices and emotionalism, the inappropriate analogies, etc., in the other fellow's argument and how to watch for and avoid the irrational in your own judgments. The author makes plain not only **how** but also **why** people resist facing the truth.

A tool for **clear thinking** as well as **convincing others.**

$9.95 *plus* **$1.25** handling

THE ART OF ARGUMENT by Giles St. Aubyn

Math Without Tears



Using non-technical language and a light touch Roy Hartkopf gives you a basic understanding of many everyday applications of mathematics. He takes the reader from simple counting to trigonometry and calculus, emphasizing the practical aspects of math. Humorously written.

Learn math in the comfort of your own home at minimum cost.

$11.95 *plus* **$1.25** handling

No handling charge on 3 books or more!

EMERSON BOOKS, INC.

Dept. 182-H Verplanck, N.Y. 10596

10-Day Money-Back Guarantee

WONDERS of the UNIVERSE 1984 CALENDAR



Featuring colorful deep-space photography from NASA and the great observatories worldwide, this high-quality calendar is a superb gift idea. It folds out to a big 11 x 22 inches and has ample space for writing notes. Important sky happenings and historic moments in space exploration highlight the days of the year.

One calendar—**$7.95** ppd. Order two or more—**$7.50** ea. ppd. (Foreign airmail postage—$5 for one. Each additional—$1.)

Please send me____ calendars for $____ postpaid.

☐ Include your complete set of catalogs **FREE** with my order.

Name____

Street____

City____

State____ Zip____

☐ MasterCard ☐ Visa
☐ Money Order ☐ Check

CREDIT CARD HOLDERS CAN ORDER TOLL-FREE 1-800-321-2369.

Expiration Date____

Card No.____

HANSEN PLANETARIUM
Dept. SA-11 15 South State Street
Salt Lake City, Utah 84111 (801) 535-7317

Speak German like a Diplomat!

What sort of people need to learn a foreign language as quickly and effectively as possible? *Foreign service personnel,* that's who.

Now you can learn to speak German with the same materials used by the U.S. State Department — the Foreign Service Institute's *Programmed Introduction to German.*

The FSI spent thousands of dollars and many years developing these materials for use by members of the United States diplomatic corps. Today people in all walks of life who need to learn to speak a foreign language are turning to this outstanding audio program.

The FSI's German Course is by far the most effective way to learn German at your own convenience and pace. It consists of a series of cassettes, recorded by native German speakers, and accompanying textbook. You simply follow the spoken and written instructions, listening and learning. By the end of the course you'll find yourself learning and speaking entirely in German!

This course turns your cassette player into a "teaching machine." With its unique "programmed" learning method, you set your own pace—testing yourself, correcting errors, reinforcing accurate responses.

AUDIO-FORUM®

This Programmed Course comes in two volumes, each shipped in a handsome library binder. Order either, or save 10% by ordering both:

☐ **Volume I. Programmed Introduction to German,** 10 cassettes (13 hr.), and 647-p. text, $125.

☐ **Volume II. Basic Course Continued,** 8 cassettes (8 hr.), and 179-p. text, $110.

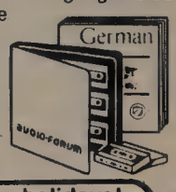
(Conn. and N.Y. residents add sales tax.)

TO ORDER BY PHONE, PLEASE CALL TOLL-FREE NUMBER: **1-800-243-1234.**

To order by mail, clip this ad and send with your name and address, and a check or money order — or charge to your credit card (AmEx, VISA, MasterCard, Diners) by enclosing card number, expiration date, and your signature.

The Foreign Service Institute's German course is unconditionally guaranteed. Try it for three weeks. If you're not convinced it's the fastest, easiest, most painless way to learn German, return it and we'll refund every penny you paid. Order Today!

112 courses in 35 other languages also available. Write us for free catalog. Our 12th year.



Audio-Forum
Suite 37N
On-the-Green,
Guilford, CT 06437
(203) 453-9794

All orders received by Dec. 12 will be delivered before the holidays!
Or visit our New York sales office 145 E. 49th St., New York, NY 10017 (212) 753-1783.

THE AMATEUR SCIENTIST

Looking into the ways of water striders, the insects that walk (and run) on water

by Jearl Walker

To me water striders are among the most fascinating of all insects. Members of the order Hemiptera, they live on the surface of the water in a slowly moving stream. Some 75 to 85 species are found in North America, and hundreds more, including many that are exclusively oceanic, are known in other parts of the world. I have been studying some of the long-legged types (of the family Gerridae), and their way of life is a marvel.

Water striders dash, leap and scramble across the water if they are disturbed, moving at a rate of about a meter per second. At other times they may move quite slowly. They do not swim. Instead they glide over the water surface by pushing horizontally against the water, much as a human sprinter pushes against starting blocks. They locate one another by means of the waves they generate as they move. They also seem to locate objects in their environment by means of waves.

Freshwater striders spend the winter in hiding under rocks or vegetation or lying on the bottom of a stream or pond. In the spring they become active, move to the surface and mate. Eggs are left by the female on submerged objects. After two weeks the nymphs hatch and move to the surface. Maturing to the adult stage takes a little more than a month.

The water striders I studied are skittish insects that apparently can see well enough to detect my approach. When I waded through streams, they were obviously also scared off by the waves I made. I learned to move slowly and then stand still for a long time so that they would eventually ignore me.

I caught several striders in a small, fine-mesh butterfly net. The long handle enabled me to avoid tipping off my presence. It was hopeless to try to chase them; they could easily outrun me in the water. When I caught one, I dropped it into a small glass jar. With this last I had to be quick; the insect could jump several centimeters in the air and out of the jar before I got the lid on.

I believe the striders I saw were all of the species *Gerris remigis*. Like all freshwater striders they had some kind of wing growth along their back. (Marine water striders are wingless.) The specimens I caught ranged in length from less than half a centimeter to about 1.5 centimeters. The immature ones were only a few millimeters long.

I rarely saw any striders along the edges of lakes, ponds or streams more than a few meters wide. On smaller streams the striders were absent from turbulent water and from places where algae covered the surface. They were also absent from regions with sparse vegetation. Areas with slowly moving water and some emergent vegetation were heavily populated with the insects. The water they seemed to like best was only a few centimeters deep. Presumably the striders congregate there because the water is too shallow for any fish large enough to eat them.

The striders were easy to detect. As soon as I waded into an appropriate habitat the water surface became a frenzy of small ripples as they scurried off. On my first explorations I found dozens of areas covered with small water striders. Then in a stream I happened on a pool busy with large ones.

For hours I sat on the shore or squatted in the water to watch their movements. Later I caught several of them and released them into a plastic container filled with water. If I took care to move slowly to avoid scaring them, they would stay on the water surface, dashing about, and would not attempt to climb over the edge.

I lay next to the container with a 20-power magnifying lens to examine how the insects manage to stand and to move on the water. Patience was called for. I waited until an insect moved into the focal plane of the lens. Then I could follow it for a short time before it became alarmed and skated off.

To photograph striders on streams I put a telephoto lens on my 35-millimeter camera, a single-lens reflex model. When I had an insect on the water in my plastic container, I hovered patiently and motionlessly with a close-up lens on the camera.

If you intend to photograph these insects in nature, you would do well to eliminate some of the light reflected from the water surface. Otherwise the photographs will be dominated by images of the surrounding trees. You can diminish the reflections by taping a polarizing filter over the lens. Mount the filter so that its axis is vertical.

All water striders have three pairs of legs. The front legs are usually short; they help to support the insect. On the striders I caught the other legs are much longer than the body. The two middle legs are primarily responsible for propulsion. The two rear legs may also aid in propulsion, but more often they serve for steering as the insect glides on the surface.

Each leg consists of several segments. From the body outward the segments are the coxa, the trochanter, the femur, the tibia and the tarsus. The tarsus itself may consist of several segments.

The long-legged water striders are distinguished by the position of a claw on the last tarsal segment, just above the end of the leg. The location of the claw may help the insect when it stands on the water, taking advantage of surface tension.

A stationary water strider rests its weight on all six legs. On the front and middle legs only the tarsus touches the water. On the rear legs the tibia and the tarsus touch. These sections are not submerged but lie in shallow depressions in the water surface. The depressions for the front legs are small. Those for the other legs are elongated because more of the leg is in contact with the water.

When I am close to a water strider with the sunlight at the right angle, the depressions are easily spotted. At other times they are apparent in the shadow the insect casts on the bottom of the stream or the container. There I see a slim shadow for the body itself and more prominent dark ovals at the end of the thin shadow of each leg. The ovals result from the bending of light rays by the curving of the water surface at the depressions.

With a flat water surface the rays of sunlight illuminating a given region would all refract into the water in the same way. Thus they would evenly illuminate the bottom of a stream or a container. When instead the rays pass through the depression surrounding a tarsus of the insect, they are sent off to the sides. The shadow cast on the bottom is therefore considerably larger than the tarsus. These large shadows proved to be useful. When the insect moved a leg, I had better luck following the oval than the leg.

Often the striders glide slowly, presumably searching for food. At other times they dart over the water. The motion is always in a straight line. At the end of a glide the insect stops, repositions one middle leg or both of them to reorient its body and then propels itself forward again.

Although I can follow the slow movements, I cannot see the details of the fast ones. In a slow movement the insect propels itself by moving its middle legs toward the rear. The rear legs are nearly stationary but might move slightly rearward too. The front legs appear to serve only as a support.

The faster movements have been recorded in high-speed motion pictures. Here the legs serve basically the same functions but the rearward motion of the middle legs lasts only about 20 milliseconds. The acceleration is large, about 10 times the acceleration of gravity. The front legs are momentarily lifted from the water surface, as is the upper end of the tibia on each rear leg. After the acceleration both sets of legs are returned to the water surface and the insect glides. The rear legs then serve as stabilizers to hold the glide straight. Eventually the kinetic energy of the movement is dissipated and the insect skids to a stop. The energy is lost to the waves generated by the movement and to the friction between the insect and the water surface.

The propulsion from the middle legs involves several rotations. The tarsus and the tibia rotate about the joint between the tibia and the femur faster than the femur rotates about its attachment to the trochanter. These two rotations push the tarsus against the rear wall of the depression in which it initially rests. The resistance of the water to the push supplies the force of propulsion.

In this motion one can see two advantages of the long middle legs of the insect. First, the length provides a long lever arm for pushing against the water surface. A human oarsman likewise finds that a long oar makes for easier rowing.

The second advantage is that the long tarsus provides more friction with the water, increasing the effectiveness of a push of the leg backward. The front legs have shorter tarsi so that the friction there is minimized, since friction shortens a glide. Some species of water strider have fans deployed from the tarsi of the middle legs to aid the push against the water. Some have projections from the legs that effectively grab the water surface, as metal studs on a snow tire bite into the snow for traction.

Much of the support for the insect as it rests or glides arises from the surface tension of the water. Surface tension is the name given to the cohesion of the water at the surface. In textbooks this phenomenon is described in terms of the electrical attraction between the water molecules. Consider one molecule that lies on the surface of pure water. Electrical forces pull it horizontally toward its neighbors, but all those forces balance out by symmetry. The molecule is also pulled downward by the neighbors just below it, but it cannot enter the bulk water because of collisions with those underlying molecules, all of which are in motion. Thus the surface is said to be in a state of tension because every molecule on it is subject to a net downward electrical force.

A second way of examining the surface is in terms of energy. Suppose the surface is to be expanded. If more surface is to be gained, molecules must be brought there from the bulk water. Such a molecule initially has no net force acting on it because on the average the forces from its neighbors balance out. At the surface, however, a net downward force develops. The movement



A water strider and its shadow



Better Than Jogging, Swimming or Cycling

NordicTrack

Jarless Total Body Cardiovascular Exerciser
Duplicates X-C Skiing For The Best Way To Fitness

NordicTrack duplicates the smooth rhythmic total body motion of XC Skiing. Recognized by health authorities as the most effective fitness building exercise available. Uniformly exercises more muscles than jogging, swimming, cycling and rowing.

Does Not cause joint or back problems as in jogging. Highly effective for weight control and muscle toning.

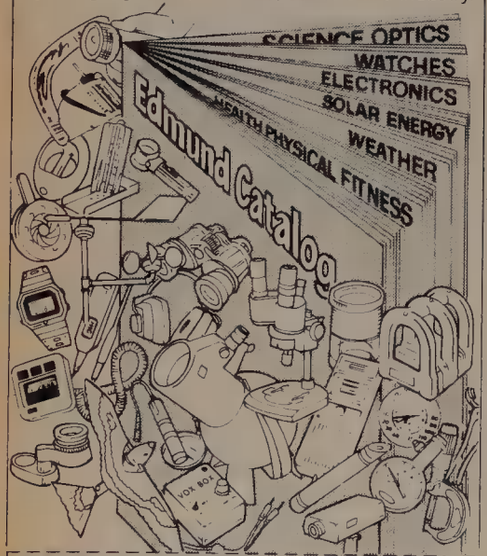
Easily Adjustable for arm resistance, leg resistance and body height. Smooth, quiet action. Folds compactly to require only 15 by 17 inches of storage area. Lifetime quality.

Used In thousands of homes and many major health clubs, universities, and corporate fitness centers.

Call or Write for FREE BROCHURE
Toll Free 1-800-328-5888 MN 612-448-6987
PSI 124 F Columbia Crt., Chaska, Minn. 55318

Send for Your Free Edmund Scientific Catalog Today!

America's largest mail-order market place of thousands of unusual and fascinating values for hobbyists, gadgeteers, students, schools and industry.

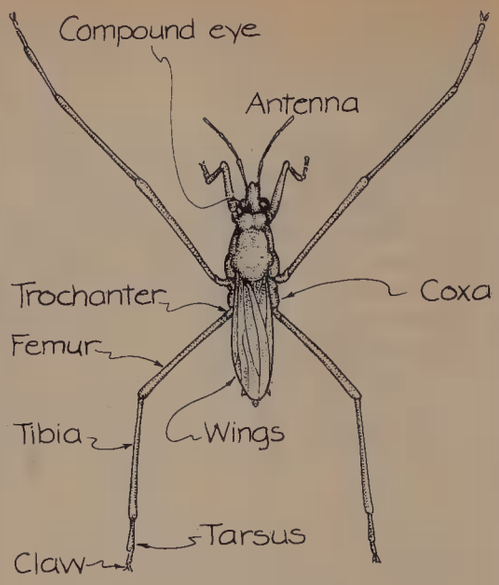


Edmund Scientific Co., Dept. S228
Edscorp Bldg., Barrington, NJ 08007
Please send me your FREE Catalog.

NAME

ADDRESS

CITY STATE ZIP



An adult strider (**Gerris remigis**)

from the bulk to the surface therefore requires energy from whatever is causing the expansion of the surface. In this description the surface is said to be in a state of tension because the water attempts to decrease its energy by contracting its surface.

The surface tension of water can be surprisingly strong. If a sewing needle is laid horizontally and carefully on water, the surface tension can hold the needle in place. The support comes from the depression of the surface produced by the weight of the needle. At first the needle starts to sink, curving the water surface, but then the surface begins to put a net upward force on the needle. Part of the upward force is due to buoyancy (because the needle has displaced some of the water). The rest of the upward force comes from the tension along the curved surface of the depression. When the demonstration is done right, the needle receives enough upward force to balance its weight, and so it floats.

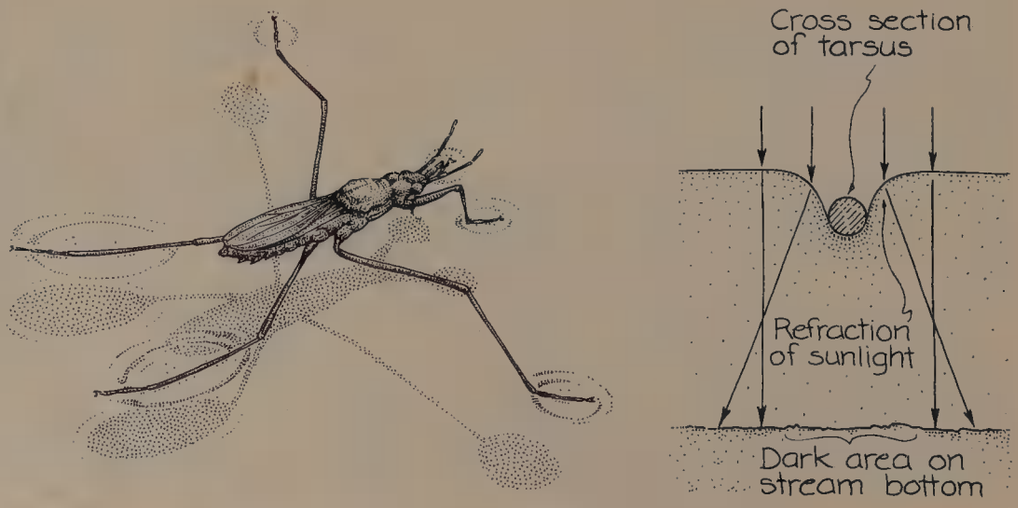
Similar support is given to an insect resting its weight on the tarsi of its legs. In addition the insect has evolved other support mechanisms. Closely spaced fine hairs on the legs and body of many species of water strider trap a layer of air. The hair, legs and body may also be coated with a waxy material to resist wetting.

The layer of trapped air provides additional buoyancy. I demonstrated the presence of air by dunking a water strider. Underwater the air on its surface shimmered in the sunlight. When I released the insect, it quickly regained the surface and scampered off, obviously still dry.

Another source of resistance to wetting has recently been discovered in the oceanic water striders. They need it because they are more exposed to rain and splashing and rarely find any floating object they can climb on to dry off. They differ from the freshwater striders in the structure of the microtrichia, the microscopic fixed hairs on the surface of the body. In the freshwater species the microtrichia are shaped like pegs and jut outward from the surface of the body. In the oceanic species they are shaped like mushrooms, presumably to aid in trapping tiny air bubbles next to the body for additional buoyancy.

Water striders do sometimes become wet, which suggests that the waxy covering or the air trapped in the hairs can fail. I observed several water striders that seemed to sink progressively into the water. The sinking strider would climb onto a rock or a lily pad to dry off.

As I was trying to get close-up photographs of one strider I provoked it into skittering across the water in my plastic container for a long time. As it began to sink it moved to the side of the container, hoisted one leg onto the plastic and hung there for a while. If it did not find the side of the container, it groomed itself dry by rubbing one front leg along the length of the adjacent middle leg, both of which were lifted out of the water. Special hairs on the tibia of the front leg serve for removing water from the other legs in such circumstances.



How the pattern of shadows from a strider is formed

The buoyancy, surface tension and resistance to wetting provide strong support for a strider when its legs are dry. The insect not only can stand on the water with all six legs but also can shift its weight to four legs while grooming. I often saw striders leaping several centimeters into the air. Although the downward push required for this maneuver created waves in the water, the insects did not break through the surface. Sometimes I saw two striders tangle in what appeared to be a fight. It always ended with one insect or both leaping into the air and then running off. Even in these leaps the insects did not break the water surface.

There is more to surface tension than I have said. Perfectly clean water is never found in nature. Even when water is thoroughly cleaned in the laboratory, its surface will be coated with a monolayer of other molecules after a few minutes of exposure to the air. Water found in nature certainly has such a top coating, which reduces the surface tension.

I believe another factor, surface viscosity, aids in the propulsion of a water strider. When the middle legs are brought rapidly to the rear, they push against the back wall of the depression they create. Because of surface viscosity the leg does not simply slide up and out of the depression in the course of the motion. It gets to give a good push. The viscosity is associated with the resistance of the top layer of molecules (the contaminating layer) to sliding over the layer of water slightly lower.

I once watched a water strider that was having difficulty in moving across a stretch of water. The water was almost stagnant and in one area was covered with a thick layer of scum that was obviously highly viscous. When I chased a strider into this area, it could no longer push off from the water and then glide. Its motion was more of a forward leap: it kicked off with its middle legs and flew through the air for several centimeters before it landed, immediately coming to a stop. After I chased it back into cleaner water it was able to glide again.

When a water strider glides, the fluid surface passing under its supporting tarsi must flow quickly into the curved shapes of the depressions. In the scummy region the high viscosity prevented this flow, but it also provided a firm support for the insect's spectacular leap.

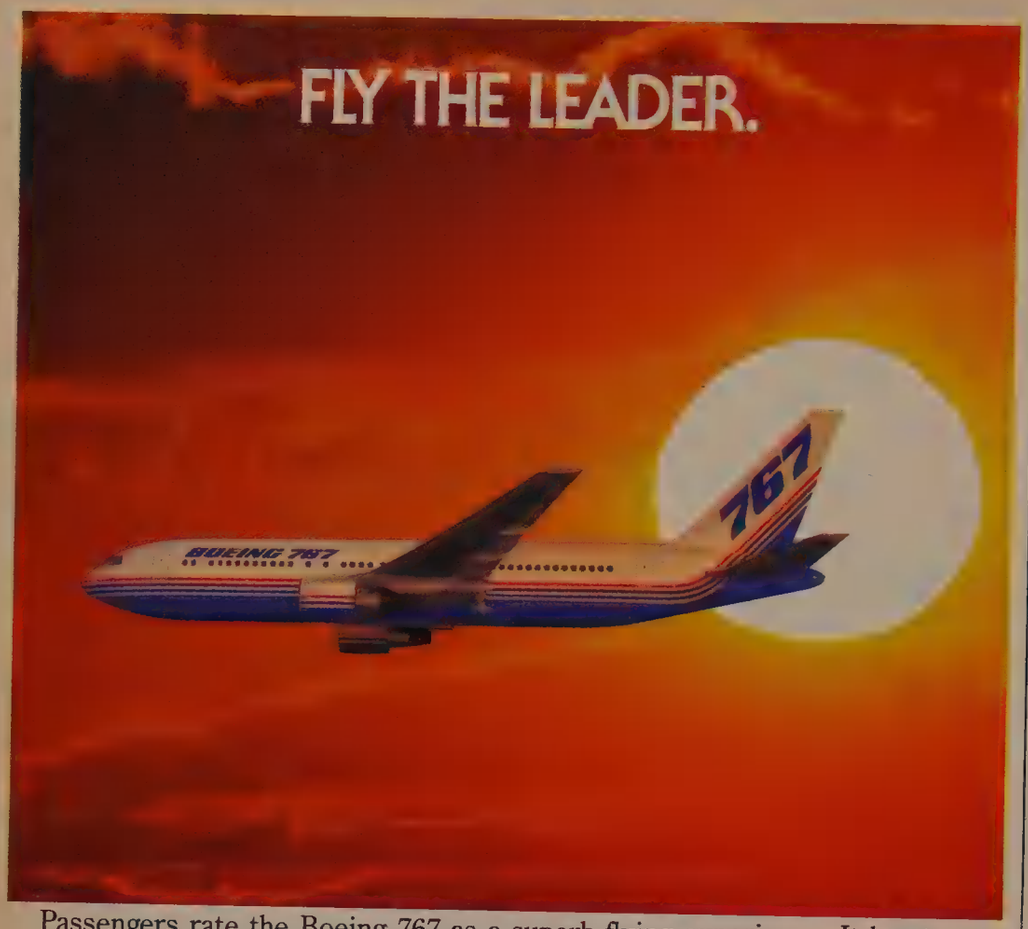
A water strider is so agile that a female can move even while mating. One afternoon I watched a pair of the insects mating for almost an hour. They stopped only when I captured them. The male had mounted the female from the rear with his weight resting on the female and on his own rear legs. The male's front legs were around the female's body just forward of her middle legs, and his middle legs were swiveled



The 757 is the most fuel-efficient jetliner in the sky. In passenger comfort, it's superior to any other plane its size. Here's the perfect match of advanced technology and passenger conveniences. It's one way Boeing is helping to keep air fares one of the world's best travel values.

BOEING
Getting people together.





Passengers rate the Boeing 767 as a superb flying experience. It has two wide aisles and is the only wide-body with seven-abreast seating, so nearly everyone has either a window or aisle seat. Pilots rate the 21st Century flight control system as the most advanced in the world. BOEING
Getting people together.



The hands that built the modern world.

In this fascinating landmark book, Landes explores the history of clocks, the people who make them, and the political, economic, and cultural importance of measurable, accurate time.

Revolution in Time "will rank high with those who delight in discovering the new or curious." —*Kirkus Reviews*

With 8 color plates, 28 b/w halftones, 13 line illustrations.

Revolution in Time

Clocks and the Making of the Modern World

David S. Landes

$20.00

The Belknap Press of

Harvard University Press

Cambridge, Massachusetts and London, England.

upward. The arrangement left the female's legs entirely unhindered. As they mated the male was motionless. The female, however, continued to move slowly over the surface. She appeared to be incapable of rapid gliding but could still move in glides of a centimeter or so.

I had supposed the ability of a water strider to stand on the water and glide over it depended critically on the wide separation of the places where the legs touched the water. At each such place the curvature of the surface in the depression around the tarsus provides an upward force on the insect. What would happen if these several depressions were brought close to one another? The surface between nearby tarsi would be flatter and would provide less upward force. Hence I supposed that a water strider would be unlikely to bring its legs close together.

My supposition was wrong. Clearly the propulsion achieved by an insect demands that the middle legs be brought close to the rear legs. For an instant as a water strider is accelerating the front legs are even lifted off the water, requiring that the middle and rear legs fully support the insect. The tarsi of those legs are then quite close to each other. Yet the insect does not break through the surface.

The mating pair I monitored had most of their weight resting on the water directly behind the female. They may have been low in the water, but they were apparently in no danger of breaking through the surface layer even when the female brought her middle legs rearward for propulsion. Furthermore, the water striders of the family Velidae have short legs, and so their support regions are closer. The insects of this family are just as agile as the long-legged striders I studied.

The wave patterns the striders create when they are moving are striking. A simple push and glide builds up a circular group of waves that expand from the site of the push. The waves arise during the acceleration stage, when the insect forces each middle leg against the rear wall of the depression in which it rests.

The full rearward motion of each leg sends out a semicircle of waves to one side. Since the two legs are synchronized, the two sets of waves merge at the front and the back of the insect. The merger is nearly perfect. When I am close to a strider, however, I can see two wakes left in the group of waves expanding to the rear. I believe they arise from the rear legs.

It is difficult to analyze the production and propagation of these waves. The waves are somehow generated when a disturbance, such as the rearward push of the strider's legs, upsets the initially flat surface of a pool of water. Two forces, operating to restore the surface

flatness, are responsible for the propagation of the waves away from the site of the disturbance.

When the wavelength of the wave is short (thus when the shape of the water surface is highly curved), surface tension is the important force. The waves are called capillary waves. For long wavelengths gravity is important, and the waves are said to be gravity waves. With intermediate situations one must consider both forces. In no instance is a wave traveling over water a simple sine wave, notwithstanding what is said in many textbooks.

The structure traveling over the surface is considered mathematically to be a superposition of many sine waves of differing wavelength. In the mathematical model they are called phase waves. Since the speed of each wave depends on its length, the waves do not travel in lockstep; hence they end up interfering with one another.

The result is often an impressive pattern. A disturbance such as an insect's kick sends out a short series of phase waves. Together they form what is called a wave group, which spreads out from the insect. On each side of the group the water is flat. Within the group some five crests can be seen. The distance between the crests is approximately the average wavelength of the phase waves making up the wave group.

The speed of the movement of the crests over the water is termed the phase speed. It differs from the speed of the wave group as a whole. When gravity dominates the propagation of the waves, the phase speed is larger than the group speed. Therefore the crests appear at the the rear of the wave group, overtake the middle and disappear at the front of the group. Since their amplitude is greatest at the middle of the group, one can follow the change in the amplitude of a crest as it traverses a wave group.

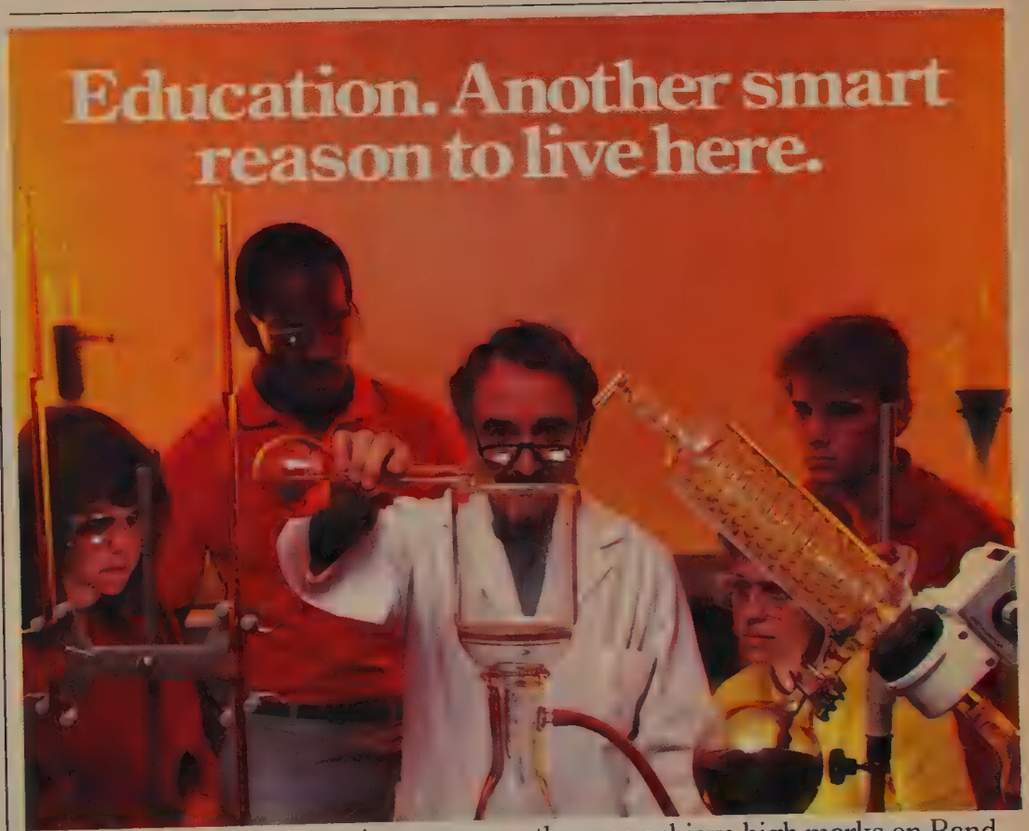
The waves generated by a water strider are capillary waves, because their wavelengths are short enough for the waves to be dominated by surface tension as the restoring force. With capillary waves the phase speed is less than the group speed. Hence a crest in an expanding wave group from a water strider appears at the front of the group, grows in amplitude as the center of the group overtakes it, decreases in amplitude as the rear of the group reaches it and finally disappears. At any instant the wave group has approximately five crests, separated by a millimeter or so.

This wave group expands roughly as a circle until it runs into some obstacle or its energy is dissipated. After gliding the insect may generate another such pattern of capillary waves. If the insect's speed relative to the water is less than .23 centimeter per second, no waves are generated. That is the minimum speed at



World travelers choose the 737 because it has one of the best on-time records. The 737-300 is the latest addition to the family. This new jetliner, with added passenger amenities, advanced flight controls and quieter engines, will make flying an even better travel experience.

BOEING
Getting people together.



In fact, education is the main reason Metro Atlanta has such a high degree of professional and skilled people. Some 36 colleges, universities and seminaries give Metro Atlanta an "A" in education. And helped the area achieve high marks on Rand McNally's report card of places to live.

ADVANTAGE METRO ATLANTA

For further information, please write: Roy Cooper, MACFED, P.O. Box 1740, Atlanta, Georgia 30301. Or call (404) 577-3838.

Clayton Cobb DeKalb Douglas Fulton Gwinnett



10 GOOD REASONS WHY EAGLE IS WORTH MORE THAN ANY ORDINARY WAGON.

The American Eagle is one luxury wagon whose value increases with every inch of snow, every drop of rain, and every mile of bad road or highway. In fact, the uglier the driving, the better Eagle looks. Because when ordinary wagons are slipping, sliding, or just plain stuck, the Eagle can switch into the sure-footed traction of 4-wheel drive to transport you confidently through winter's worst storms.

But a feeling of security is by no means Eagle's only luxury. The interior room and comfort is outstanding. And with standard features like power steering and brakes, plush 18 oz. carpeting, woodgrain instrument panel overlay and individual reclining seats of genuine leather, the Eagle Limited makes 4-wheel drive extremely attractive.

It's a great looking wagon, but more importantly, it's rugged, eager and very reliable. No wonder the National Ski Patrol has chosen the Eagle as its official vehicle. They *have* to get through.

NATIONAL SKI PATROL

In addition to impressing you with its 4-wheel drive performance, Eagle can switch back to 2-wheel drive to impress you with its outstanding gas mileage, (24) EPA EST. MPG and 30 HWY. EST.*

The Eagle is the perfect all-season, all-weather family wagon. Because when other luxury wagons are grounded, the Eagle flies.

SAFETY IS A SNAP



THE 2-WHEEL/4-WHEEL DRIVE Eagle

AMERICAN MOTORS

* 4-cylinder/4-speed. Use these figures for comparison. Your results may differ due to driving speed, weather conditions and trip length. Actual highway figures are probably lower.

which waves can propagate over the surface of water. Insects moving at less than this speed leave the water undisturbed except for the wave group initiated by their propulsion.

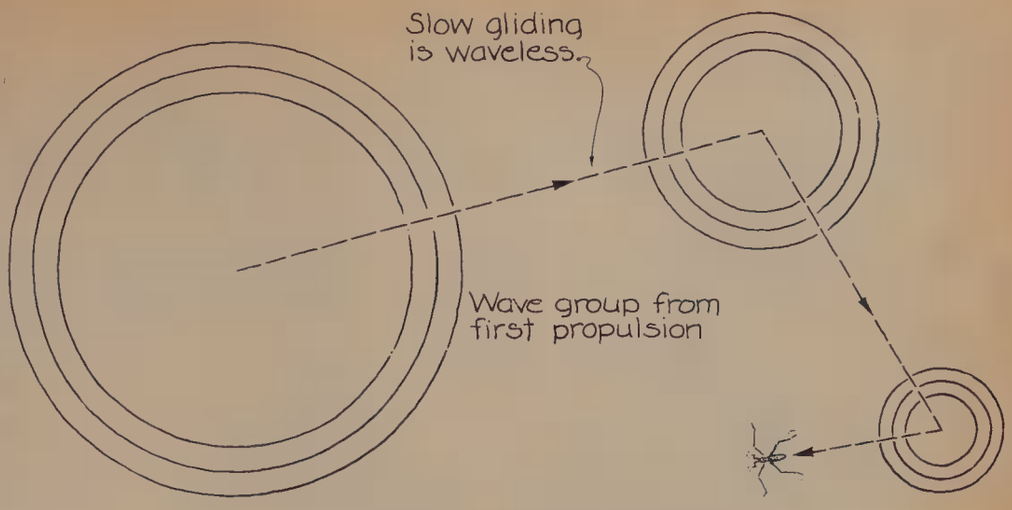
If the insect glides faster than the minimum wave speed, it continuously generates waves. The pattern that results has two curious features. One feature is that the waves in front of the insect are more closely spaced than the ones in the rear. The second feature is that the wave pattern takes the shape of a *V*, with the apex at the insect's head.

In 1883 Lord Rayleigh demonstrated mathematically that a small object moving in water with a relative speed in excess of .23 centimeter per second sends out waves shorter in length toward the front than at the rear. He was not concerned with wave patterns from insects but with the pattern produced by a fishing line in a moving stream. The line served as an obstacle to the flow of water. (Although the line was stationary and the insect moves, the difference is irrelevant. It is the relative speed of the object and the water that matters.) The generation of waves by the obstacle builds up closely spaced wave crests upstream of the obstacle. They are parts of capillary waves. Downstream the perturbation of the water by the obstacle sends out widely spaced wave crests. They are parts of gravity waves.

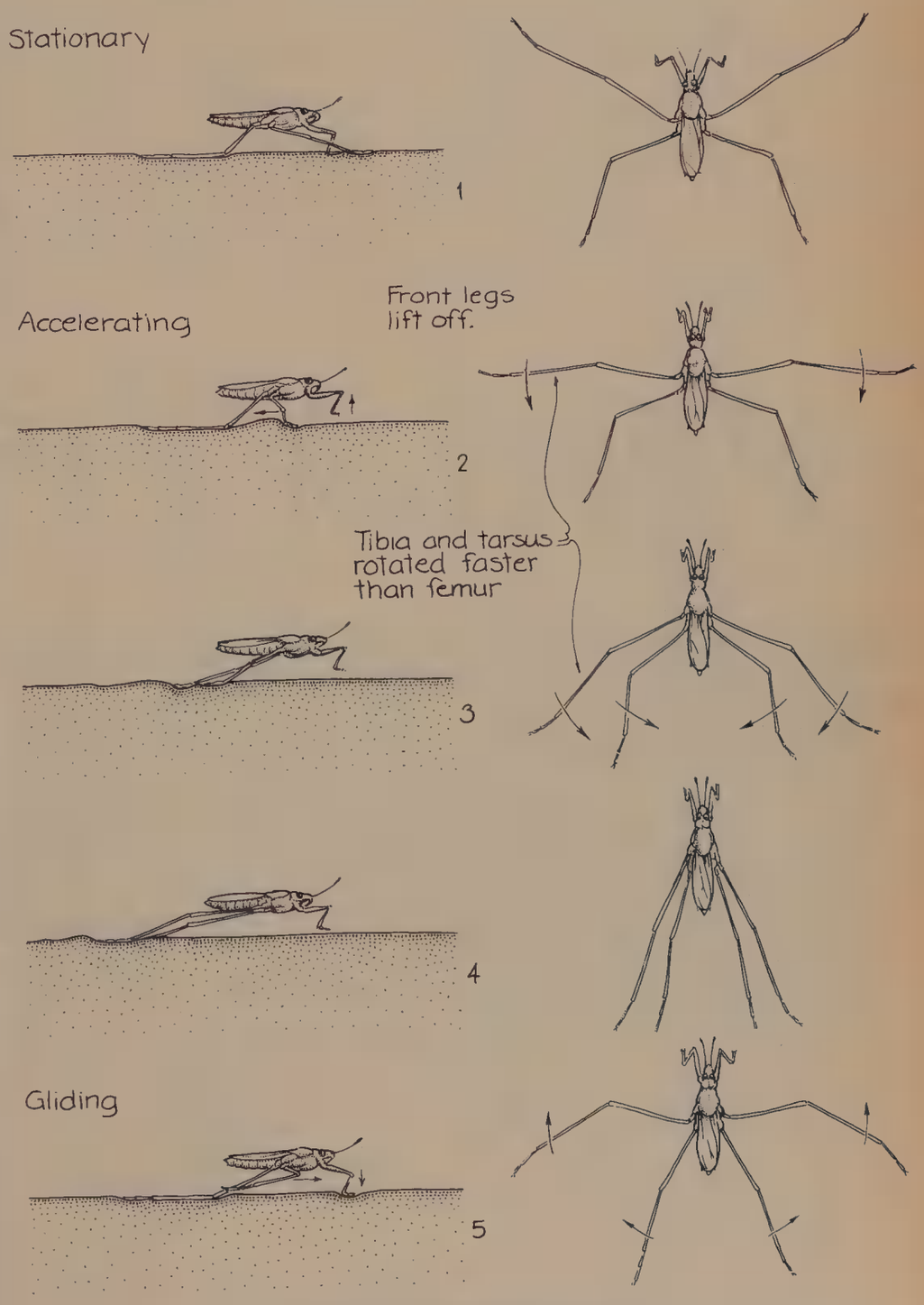
In the *V* wave pattern around a water strider the capillary waves are outside the *V*, the gravity waves inside. The angle of the *V* reflects a mathematical relation of twice the minimum wave speed of .23 centimeter per second divided by the relative speed of the obstacle and the water. A water strider in a faster glide ends up generating a narrower *V* wake.

In 1972 R. Stimson Wilcox of the Australian National University demonstrated that the surface waves generated by water striders play an important role in their mating. He studied the genus *Rhagadotarsus* Breddin by capturing several specimens and installing them in a large tray of water. The male initiated mating by generating wave signals. It gripped a support in the water with its front legs while its middle legs (and possibly the rear ones too) created the waves.

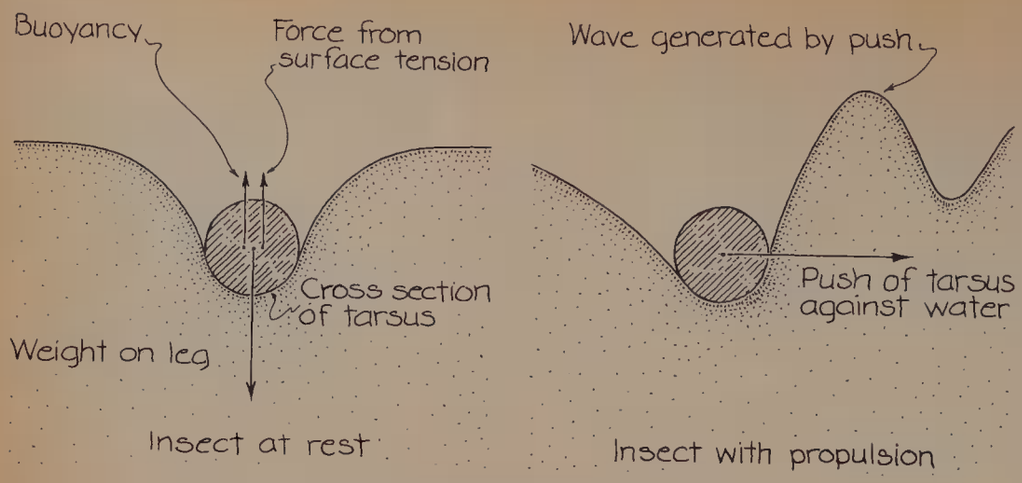
Stimson discovered three main types of signal preceding copulation, all with a frequency of between 17 and 29 waves per second. The male could call for a female by sending out a group of from seven to 15 waves at high amplitude. Alternatively it could transmit a group of only two or three waves in what Stimson interpreted as more a signal for courtship. If a female approached, the signal switched rapidly to a strictly courtship mode of 30 waves at low amplitude. When the female was within a few centimeters of the male, the female



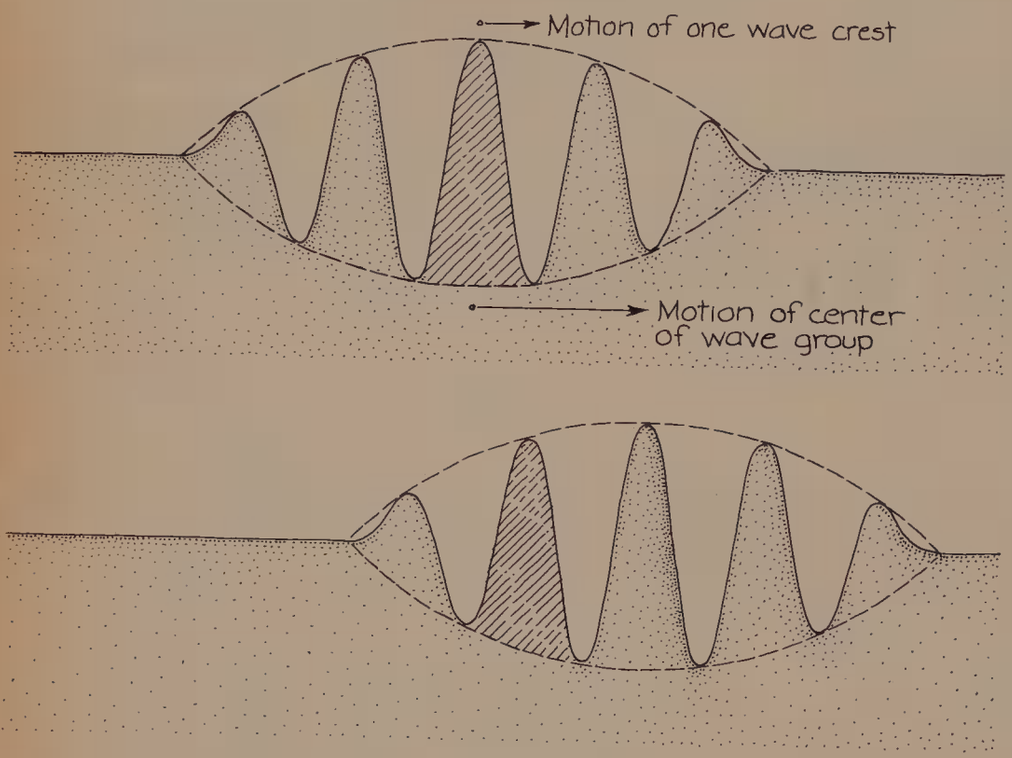
The skitter path of a waterborne insect



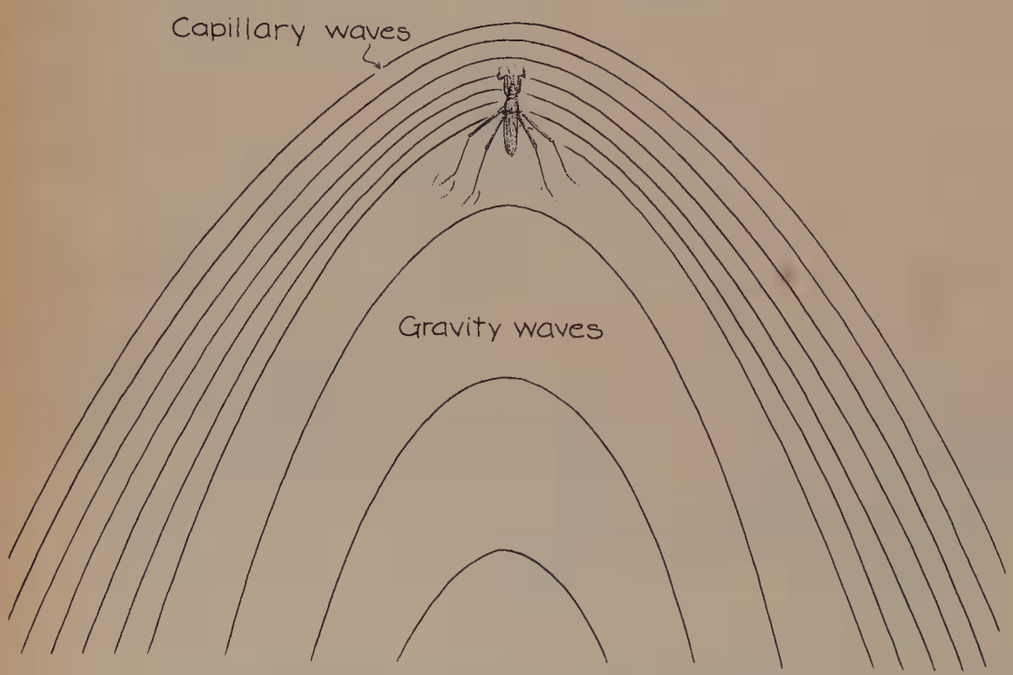
Stages in the movement of a strider



The interaction of a strider's leg with the water surface



The motion of a capillary-wave group



The pattern of waves around a rapidly moving strider

too would send such a signal. Copulation then began.

A male sometimes generated another signal of higher frequency in an apparent warning to other males to avoid the mating area. If two males closed on each other, they would fight, perhaps for some minutes. I saw several such fights as water striders circled and then leaped at each other, after which one would chase the other.

On all my trips up and down streams I found water striders apparently locating one another by means of the waves they generated. When a wave group reached a water strider, the insect paused and then oriented itself perpendicular to the wave crests and toward the source of the crests. Then it dashed a short distance toward the source. After waiting for a fresh set of waves to pass, it would again dash off in the direction of the source. I cannot believe the insect located the source entirely by sight; if it did, the occasional pauses would be unnecessary.

Striders also employ waves to find prey. When a fly fell into the water and flapped violently on the surface, a water strider located it by means of the resulting waves. Sight might have played a role when the strider got near the fly, but not before.

Although the strider's exploitation of waves seemed obvious to me, I wanted a better test. I tried oscillating a twig in a stream, but only one strider came to investigate. It might have arrived there even if I had not been dabbling in the water. I tried a vibrator sold for massage. The oscillating part, which is fully encased in rubber, is powered through a wire running to two small batteries. (A vibrator run on household current would invite electrocution.) A switch on the battery holder varied the frequency of oscillation.

I waded into the center of a pool of water about six meters wide along a stream. Several water striders collected at the edges of the pool after sensing my arrival. I submerged part of the vibrator and set its oscillation at about 20 hertz. The striders immediately paused in their motion, turned toward me and within about three seconds raced to within a few centimeters of the vibrator. Neither the vibrator nor I looked like a water strider, and so the insects surely were attracted by the waves.

Insects of many other types live on or near the water surface in a stream. You might like to search for whirligig beetles. The capillary waves they generate were studied by Vance A. Tucker of Duke University, whose work provided the basis for my discussion of the waves from a water strider. Another interesting aquatic subject is the back swimmer. This insect travels upside down and just below the surface of the water by rowing with its legs.

HOW TO REACH THE PEOPLE WHO MAKE THE FUTURE HAPPEN IN SPAIN

INVESTIGACION Y CIENCIA is the Spanish edition of SCIENTIFIC AMERICAN which attracts a young audience of affluent professionals who have come to the top because of their technical expertise. Almost half of them hold top management job titles. The fortunes of major corporations in Spain increasingly depend on technically sophisticated people. INVESTIGACION Y CIENCIA keeps these people in touch with the advances in science that drive the growth of Spain's industry.

SCIENTIFIC AMERICAN speaks the languages of more than half the world's population. We are the one publication in the world today providing efficient coverage of technology-based management. We reach men and women in industry, whose qualification to make technical decisions places them in key positions in their country's government and industry.

SCIENTIFIC AMERICAN in eight languages: English, Spanish, Italian, French, German, Japanese, Chinese and now Russian has gathered in its audience three million people who make the future happen around the world.



For more information on our Spanish-language edition contact:

Elena Sanchez-Fabres
Prensa Scientifica, s.a.
Calabria 235-239
Barcelona 29, SPAIN
Telephone (011) 343-322-0551

or in New York
John Kirby,
V.P./Advertising Director
SCIENTIFIC AMERICAN
415 Madison Avenue
New York, New York 10017
Telephone 212-754-0262

SAY "CELESTRON"

NOVEMBER 11th AT YOUR CELESTRON DEALER

Free telescope demonstrations, free gifts, free rebates!

If you've ever looked at the night sky with wonder, November 11th is your day and night!

First the entertainment. All day on November 11th, and until closing, participating Celestron Dealers will be featuring exciting demonstrations and displays of a wide variety of Celestron telescopes — including our new inexpensive Cometron refractors, our new quality Celestron refractors and our legendary Cassegrain folded-optics telescopes. Many dealers will feature our new Super C8 — an upgraded version of the world's best selling modern telescope. There will be experts available for questions, fun and festivities. And at night on the 11th many dealers will set up scopes outside their stores — if you've never seen the wonders of the universe, this is your night. And it's all free.

Free gift. Just for walking in and saying "Celestron," participating Dealers will give you a free moon map.

Finally, the rebates. You don't have to wait till Nov. 11th, the money's right here. For the first time ever, a $10 to $20 rebate on Celestron's precision binoculars — including zoom, armoured and waterproof models and the NOVA and GIANT series. Plus a chance at every participating dealership to win a pair of Celestron 8 x 21 binoculars. Mark your calendar. And call our toll-free number.



$$ REBATES! $$



To celebrate, Celestron is offering up to $20 rebates on our full line of quality binoculars. All models are included — zoom, armoured, waterproof, mini's, Nova's and Giant series. Offer ends 12/31/83. At participating dealers.

CELESTRON®
PRECISION OPTICS

For the Names of Participating Dealers near You

CALL FREE
1-800-421-1526
(Mon-Fri: 8-4:00 Pacific Time: Continental USA only)

CELESTRON INTERNATIONAL, P.O. Box 3578 SA11
2835 Columbia St., Torrance, CA 90503 (213) 328-9560

BIBLIOGRAPHY

Readers interested in further explanation of the subjects covered by the articles in this issue may find the following lists of publications helpful.

COMPUTER RECREATIONS

SCIENTIFIC AND ENGINEERING PROBLEM-SOLVING WITH THE COMPUTER. William Ralph Bennett, Jr. Prentice-Hall, Inc., 1976.

HOW ARTIFICIAL IS INTELLIGENCE? W. R. Bennett, Jr., in *American Scientist,* Vol. 65, No. 6, pages 694–702; November–December, 1977.

THE UNCERTAINTIES OF A PREEMPTIVE NUCLEAR ATTACK

WHY ICBMS CAN SURVIVE A NUCLEAR ATTACK. Lt. Col. Joseph J. McGlinchey and Dr. Jakob W. Seelig in *Air Force Magazine,* Vol. 57, No. 9, pages 82–85; September, 1974.

STRATEGIC VULNERABILITY: THE BALANCE BETWEEN PRUDENCE AND PARANOIA. John D. Steinbruner and Thomas M. Garwin in *International Security,* Vol. 1, No. 1, pages 138–181; Summer, 1976.

BALLISTIC MISSILE GUIDANCE AND TECHNICAL UNCERTAINTIES IN COUNTERSILO ATTACKS. Matthew Bunn and Kosta Tsipis. Program in Science and Technology for International Security, Report No. 9. Massachusetts Institute of Technology, July, 1983.

SLURRY PIPELINES

THE LAMINAR-TURBULENT TRANSITION FOR FLOW IN PIPES, CONCENTRIC ANNULI, AND PARALLEL PLATES. Richard W. Hanks in *A.I.Ch.E. Journal,* Vol. 9, No. 1, pages 45–48; January, 1963.

CROSS COUNTRY COAL PIPE LINE HYDRAULICS. E. J. Wasp, T. J. Regan, J. Withers, P. A. C. Cook and J. T. Clancey in *Pipe Line News,* Vol. 35, No. 7, pages 20–28; July, 1963.

SLURRY PIPING SYSTEMS: TRENDS, DESIGN METHODS, GUIDELINES. Thomas C. Aude, Norman T. Cowper, Terry C. Thompson and Edward J. Wasp in *Chemical Engineering,* Vol. 78, No. 15, pages 74–90; June 28, 1971.

CENTAURUS A: THE NEAREST ACTIVE GALAXY

ACTIVE GALACTIC NUCLEI. Edited by C. Hazard and S. Mitton. Cambridge University Press, 1979.

THE X-RAY STRUCTURE OF CENTAURUS A. E. D. Feigelson, E. J. Schreier, J. P. Delvaille, R. Giacconi, J. E. Grindlay and A. P. Lightman in *The Astrophysical Journal,* Vol. 251, No. 1, Part 1, pages 31–51; December 1, 1981.

EXTRAGALACTIC RADIO SOURCES. Edited by David S. Heeschen and Campbell M. Wade. D. Reidel Publishing Co., 1982.

THE X-RAY JETS OF CENTAURUS A AND M87. Eric D. Feigelson and Ethan J. Schreier in *Sky and Telescope,* Vol. 65, No. 1, pages 6–12; January, 1983.

A MOLECULAR BASIS OF CANCER

PASSAGE OF PHENOTYPES OF CHEMICALLY TRANSFORMED CELLS VIA TRANSFECTION OF DNA AND CHROMATIN. Chiaho Shih, Ben-Zion Shilo, Mitchell P. Goldfarb, Ann Dannenberg and Robert A. Weinberg in *Proceedings of the National Academy of Sciences of the United States of America,* Vol. 76, No. 11, pages 5714–5718; November, 1979.

ONCOGENES OF SPONTANEOUS AND CHEMICALLY INDUCED TUMORS. R. A. Weinberg in *Advances in Cancer Research,* Vol. 36, pages 149–163; 1982.

ONCOGENES. J. Michael Bishop in *Scientific American,* Vol. 246, No. 3, pages 80–92; March, 1982.

SOLID-STATE SUPERLATTICES

SUPERLATTICE AND NEGATIVE DIFFERENTIAL CONDUCTIVITY IN SEMICONDUCTORS. L. Esaki and R. Tsu in *IBM Journal of Research and Development,* Vol. 14, No. 1, pages 61–65; January, 1970.

ELECTRON STATES IN CRYSTALS WITH "NIPI-SUPERSTRUCTURE." G. H. Döhler in *Physica Status Solidi (b),* Vol. 52, No. 1, pages 79–92; July 1, 1972.

ELECTRICAL AND OPTICAL PROPERTIES OF CRYSTALS WITH "NIPI-SUPERSTRUCTURE." G. H. Döhler in *Physica Status Solidi (b),* Vol. 52, No. 2, pages 533–545; August 1, 1972.

CONFINED CARRIER QUANTUM STATES IN ULTRATHIN SEMICONDUCTOR HETEROSTRUCTURES. Raymond Dingle in *Festkörperprobleme/Advances in Solid State Physics,* Vol. 15, pages 21–48, 1975.

ELECTRONIC PROPERTIES OF TWO-DIMENSIONAL SYSTEMS: PROCEEDINGS OF THE INTERNATIONAL CONFERENCE ON THE ELECTRONIC PROPERTIES OF TWO-DIMENSIONAL SYSTEMS. In *Surface Science,* Vol. 73; May, 1978.

INAS-GASB SUPERLATTICES—SYNTHESIZED SEMICONDUCTORS AND SEMIMETALS. L. Esaki in *Journal of Crystal Growth,* Vol. 52, Part 1, pages 227–240; April, 1981.

ELECTRONIC PROPERTIES OF TWO-DIMENSIONAL SYSTEMS: PROCEEDINGS OF THE FOURTH INTERNATIONAL CONFERENCE ON ELECTRONIC PROPERTIES OF TWO-DIMENSIONAL SYSTEMS. In *Surface Science,* Vol. 113, Nos. 1–3; January, 1982.

THE SOCIAL ARCHAEOLOGY OF MEGALITHIC MONUMENTS

BEFORE CIVILISATION: THE RADIOCARBON REVOLUTION AND PREHISTORIC EUROPE. Colin Renfrew. Cambridge University Press, 1979.

INVESTIGATIONS IN ORKNEY. Colin Renfrew. Society of Antiquaries, distributed by Thames and Hudson, 1979.

MEGALITHIC MONUMENTS. Glyn Daniel in *Scientific American,* Vol. 243, No. 1, pages 78–90; July, 1980.

A SEED-EATING BEETLE'S ADAPTATIONS TO A POISONOUS SEED

CHEMICAL ECOLOGY. Ernest Sondheimer and John B. Simeone. Academic Press, 1970.

INTRODUCTION TO ECOLOGICAL BIOCHEMISTRY. J. B. Harborne. Academic Press, 1982.

THE NON-PROTEIN AMINO AND IMINO ACIDS: BIOLOGICAL, BIOCHEMICAL AND TOXOLOGICAL PROPERTIES. Gerald A. Rosenthal. Academic Press, 1982.

THE ERUPTION OF KRAKATAU

THE KRAKATOA ERUPTION. R. D. M. Verbeek in *Nature,* Vol. 30, No. 757, pages 10–15; May 1, 1884.

THE ERUPTION OF KRAKATOA AND SUBSEQUENT PHENOMENA. Edited by G. J. Symons. Trübner & Co., 1888.

THE KRAKATOA AIR-SEA WAVES: AN EXAMPLE OF PULSE PROPAGATION IN COUPLED SYSTEMS. David Harkrider and Frank Press in *The Geophysical Journal of the Royal Astronomical Society,* Vol. 13, Nos. 1–3, pages 149–159; July, 1967.

TSUNAMIS OF VOLCANIC ORIGIN: SUMMARY OF CAUSES, WITH PARTICULAR REFERENCE TO KRAKATOA, 1883. J. H. Latter in *Bulletin Volcanologique,* Vol. 44, No. 3, pages 467–490; 1981.

THE 1883 ERUPTION OF KRAKATAU. Stephen Self and Michael R. Rampino in *Nature,* Vol. 294, No. 5843, pages 699–704; December 24/31, 1981.

THE AMATEUR SCIENTIST

WAVE-MAKING BY WHIRLIGIG BEETLES (GYRINIDAE). Vance A. Tucker in *Science,* Vol. 166, No. 3907, pages 897–899; November 14, 1969.

COMMUNICATION BY SURFACE WAVES: MATING BEHAVIOR OF A WATER STRIDER (GERRIDAE). R. Stimson Wilcox in *Journal of Comparative Physiology,* Vol. 80, pages 255–266; 1972.

Marygrove College Library
Detroit, Michigan 48221
PLEASE DO NOT REMOVE

SCIENTIFIC AMERICAN



HUMAN-POWER LAND VEHICLES

$2.50

December 1983



DODGE SERIES 600. FRONT-WHEEL DRIVE, ELECTRONIC SOPHISTICATION, 5/50 PROTECTION AND A CHOICE. The Dodge Series 600. A sporty convertible. A touring sedan. A family sedan. Even a coupe. Four distinctly unique automobiles designed for distinctly unique drivers. But it's what they have in common that makes them truly unique.



600 ES INTERIOR

All have front-wheel drive for alert handling on slippery surfaces. All possess a lean, aerodynamic design to enhance fuel efficiency. All come equipped with over 35 standard features. All can be ordered with turbo power. All are well-endowed with electronic sophistication: a computer which constantly monitors and fine-tunes the engine for optimum performance and efficiency; even an available voice alert that keeps you informed with 11 different messages. And all are backed by the industry's only standard 5/50 Protection Plan.†

As for mileage, the 600 ES—a performance sedan—actually rates an impressive 39 est. hwy. and 24 EPA est. mpg.*

Ask your Dodge dealer about buying or leasing† one of the new Dodge 600s. The revolutionary things they have in common make them uncommon cars and uncommon values.

The New Chrysler Technology. Quality backed by 5/50 Protection.

†5/50 Protection Plan. 5 years 50,000 miles, whichever comes first. Limited warranty on engine, powertrain, outer body rust-through. A deductible applies. Leases excluded. SEE DEALER FOR DETAILS. *Use EPA est. mpg

Established 1845 **SCIENTIFIC AMERICAN** December 1983 Volume 249 Number 6

ARTICLES

49 **MODERN ICEBREAKERS, by John D. Harbron**
The demand for year-round commerce at high latitudes has given rise to unusual new ship designs.

56 **IO, by Torrence V. Johnson and Laurence A. Soderblom**
One of the four Galilean moons of Jupiter is volcanically the most active body in the solar system.

68 **AN EARLY IRON AGE FARM COMMUNITY IN CENTRAL EUROPE, by Peter S. Wells** Excavations in Bavaria yield clues to an economy of the period from 1000 to 800 B.C.

94 **THE DNA HELIX AND HOW IT IS READ, by Richard E. Dickerson**
X-ray analysis of DNA shows how the architecture of the helical molecule influences its reading.

112 **QUANTUM GRAVITY, by Bryce S. DeWitt**
In a quantum-mechanical theory of gravity the very geometry of space and time would fluctuate.

130 **PREY SWITCHING IN A SIMPLE ECOSYSTEM, by Arthur T. Bergerud**
The decline of the caribou in Newfoundland was a clue to a complex cycle of predators and prey.

142 **THE AERODYNAMICS OF HUMAN-POWERED LAND VEHICLES, by Albert C. Gross, Chester R. Kyle and Douglas J. Malewicki** The right bicycle can go 60 m.p.h.

154 **THE INTERPRETATION OF VISUAL ILLUSIONS, by Donald D. Hoffman**
The visual system interprets ambiguous images according to unconscious rules of inference.

DEPARTMENTS

6 LETTERS
11 50 AND 100 YEARS AGO
14 THE AUTHORS
19 COMPUTER RECREATIONS
30 BOOKS
76 SCIENCE AND THE CITIZEN
164 THE AMATEUR SCIENTIST
172 ANNUAL INDEX
178 BIBLIOGRAPHY

BOARD OF EDITORS Gerard Piel (Publisher), Dennis Flanagan (Editor), Brian P. Hayes (Associate Editor), Philip Morrison (Book Editor), John M. Benditt, Peter G. Brown, Michael Feirtag, Robert Kunzig, Diana Lutz, Jonathan B. Piel, John Purcell, James T. Rogers, Armand Schwab, Jr., Joseph Wisnovsky

ART DEPARTMENT Samuel L. Howard (Art Director), Steven R. Black (Assistant Art Director), Ilil Arbel, Edward Bell

PRODUCTION DEPARTMENT Richard Sasso (Production Manager), Carol Eisler and Leo J. Petruzzi (Assistants to the Production Manager), Carol Hansen (Electronic Composition Manager), Carol Albert, Karen Friedman, Lori Mogol, Karen O'Connor, Julio E. Xavier

COPY DEPARTMENT Sally Porter Jenks (Copy Chief), Debra Q. Bennett, Mary Knight, Dorothy R. Patterson

GENERAL MANAGER George S. Conn

ADVERTISING DIRECTOR C. John Kirby

CIRCULATION MANAGER William H. Yokel

SECRETARY Arlene Wright

SCIENTIFIC AMERICAN (ISSN 0036-8733), PUBLISHED MONTHLY BY SCIENTIFIC AMERICAN, INC., 415 MADISON AVENUE, NEW YORK, N.Y. 10017. COPYRIGHT © 1983 BY SCIENTIFIC AMERICAN, INC. ALL RIGHTS RESERVED. PRINTED IN THE U.S.A. NO PART OF THIS ISSUE MAY BE REPRODUCED BY ANY MECHANICAL, PHOTOGRAPHIC OR ELECTRONIC PROCESS, OR IN THE FORM OF A PHONOGRAPHIC RECORDING, NOR MAY IT BE STORED IN A RETRIEVAL SYSTEM, TRANSMITTED OR OTHERWISE COPIED FOR PUBLIC OR PRIVATE USE WITHOUT WRITTEN PERMISSION OF THE PUBLISHER. SECOND-CLASS POSTAGE PAID AT NEW YORK, N.Y., AND AT ADDITIONAL MAILING OFFICES. AUTHORIZED AS SECOND-CLASS MAIL BY THE POST OFFICE DEPARTMENT, OTTAWA, CANADA, AND FOR PAYMENT OF POSTAGE IN CASH. SUBSCRIPTION RATE: $24 PER YEAR IN THE U.S. AND ITS POSSESSIONS; $33 PER YEAR IN ALL OTHER COUNTRIES. POSTMASTER: SEND ADDRESS CHANGES TO SCIENTIFIC AMERICAN, 415 MADISON AVENUE, NEW YORK, N.Y. 10017.

SCIENTIFIC AMERICAN

CORRESPONDENCE

Offprints of more than 1,000 selected articles from earlier issues of this magazine, listed in an annual catalogue, are available at $1.25 each. Correspondence, orders and requests for the catalogue should be addressed to W. H. Freeman and Company, 4419 West 1980 South, Salt Lake City, UT 84121. Offprints adopted for classroom use may be ordered direct or through a college bookstore. Sets of 10 or more Offprints are collated by the publisher and are delivered as sets to bookstores.

Photocopying rights are hereby granted by Scientific American, Inc., to libraries and others registered with the Copyright Clearance Center (CCC) to photocopy articles in this issue of SCIENTIFIC AMERICAN for the flat fee of $2 per copy of each article or any part thereof. Such clearance does not extend to the photocopying of articles for promotion or other commercial purposes. Correspondence and payment should be addressed to Copyright Clearance Center, Inc., 21 Congress Street, Salem, MA 01970. Specify CCC Reference Number ISSN 0036–8733/83. $2.00 + 0.00.

Editorial correspondence should be addressed to The Editors, SCIENTIFIC AMERICAN, 415 Madison Avenue, New York, NY 10017. Manuscripts are submitted at the authors' risk and will not be returned unless they are accompanied by postage.

Advertising correspondence should be addressed to C. John Kirby, Advertising Director, SCIENTIFIC AMERICAN, 415 Madison Avenue, New York, NY 10017.

Subscription correspondence should be addressed to Subscription Manager, SCIENTIFIC AMERICAN, P.O. Box 5969, New York, NY 10017. The date of the last issue on your subscription is shown in the upper right-hand corner of each month's mailing label. For change of address notify us at least four weeks in advance. Please send your old address (if convenient, on a mailing label of a recent issue) as well as the new one.

Name

New Address

Street

City

State and ZIP

Old Address

Street

City

State and ZIP



THE COVER

The painting on the cover shows the Vector Single, a fully streamlined tricycle that the rider pedals in a supine position, that is, while essentially lying on his back. The vehicle embodies a number of advances that have been made in recent years in the aerodynamics of bicycles and other pedaled machines (see "The Aerodynamics of Human-powered Land Vehicles," by Albert C. Gross, Chester R. Kyle and Douglas J. Malewicki, page 142). The Vector Single was designed and built by a team headed by Allan A. Voigt, president of Versatron Research, Inc. According to Voigt, the tricycle is theoretically capable of attaining a speed of almost 62 miles per hour on level ground with an input of one horsepower from the rider. A companion vehicle, the Vector Tandem, pedaled by two riders positioned back to back, holds the speed record of 62.92 m.p.h., which was achieved with an input of slightly more than 1 h.p. from each rider. It averaged 50.5 m.p.h. for 40 miles on an interstate highway.

THE ILLUSTRATIONS

Cover painting by Ted Lodigensky

Page	Source
20–28	Ilil Arbel
50	Wärtsilä
51–55	Walken Graphics
57–63	National Aeronautics and Space Administration, U.S. Geological Survey
64–66	Ian Worpole
67	John T. Trauger, California Institute of Technology
69	Caroline Quillian Stubbs
70–74	Alan D. Iselin
95	Nelson Max, Lawrence Livermore National Laboratory
96–100	Irving Geis
101	Richard E. Dickerson, Molecular Biology Institute, University of California at Los Angeles
102	Irving Geis
103	Richard E. Dickerson, Molecular Biology Institute, University of California at Los Angeles
104	Irving Geis
105	Richard E. Dickerson, Molecular Biology Institute, University of California at Los Angeles
109–111	Irving Geis
113–129	George V. Kelvin
131–140	Tom Prentiss
143	Bruce A. Sargeant
144–148	Walken Graphics
149–151	Albert E. Miller
152	Walken Graphics
154–162	Gabor Kiss
165–166	Frode Wissing
167–170	Michael Goodman



Every year at about this time, the world lights a candle to Peace.

It is a universal gesture, immune to politics. It pays tribute to no narrow philosophy, materialistic or messianic.

The candle when lit by men of good will serves the good that is in us all; it cauterizes all that in us is evil.

To every force arrayed in anger against an opposing force in the world, the candle's message is simple:

"The sum of the good you share in your hearts is far greater than those evils that seek to divide you."

The candle, when ignited, does not detonate. It illuminates. It doesn't challenge the heavens in a thundering roar. It makes its plea in a steady flame, reflected in the eye of a wondering child, or the squint of an elder who has seen it all.

The candle graces the menorah as it fits the tree. It is at home in all the windows of the world. Its reassurance cannot be measured.

For, the good news is that Peace is not to be found in the frantic mathematics of the megaton.

It reposes, calmly, in the Magic of the candle.

Light a candle to Peace this season.

LANDS' END
DIRECT MERCHANTS

If you wish a copy of this message, please write to:
Lands' End, Direct Merchants
Dodgeville, Wisconsin 53595

Play the finest

HARPSICHORDS & FORTEPIANOS

Build it yourself from one of our kits or let our experienced craftsmen build it for you.

Three Centuries of Harpsichord Making by Frank Hubbard (Harvard Univ. Press. 397 pgs. $20.00 postpaid)



For brochure send $1.00 to:
FRANK HUBBARD INCORPORATED
144-S Moody St., Waltham, MA 02154
(617) 894-3238



Better Than Jogging, Swimming or Cycling

NordicTrack

Jarless Total Body Cardiovascular Exerciser
Duplicates X-C Skiing For The Best Way To Fitness

NordicTrack duplicates the smooth rhythmic total body motion of XC Skiing. Recognized by health authorities as the most effective fitness building exercise available. Uniformly exercises more muscles than jogging, swimming, cycling and rowing.

Does Not cause joint or back problems as in jogging. Highly effective for weight control and muscle toning.

Easily Adjustable for arm resistance, leg resistance and body height. Smooth, quiet action. Folds compactly to require only 15 by 17 inches of storage area. Lifetime quality.

Used In thousands of homes and many major health clubs, universities, and corporate fitness centers.

Call or Write for FREE BROCHURE
Toll Free 1-800-328-5888 MN 612-448-6987
PSI 124 F Columbia Crt., Chaska, Minn. 55318

LETTERS

Sirs:

Derek Bickerton, in his article "Creole Languages" [SCIENTIFIC AMERICAN, July], argues very convincingly for what is undoubtedly the most important idea in modern linguistic theory: the innate knowledge in human beings of the structure of language. Creole languages are without question evidence of the strongest kind for this position, since, as Bickerton argues, how else are we to explain the strong structural similarities between all creoles no matter what their linguistic and cultural derivation? Bickerton's discussion of child language is also most suggestive: it seems that children are in some sense expecting to encounter a creole grammar when they begin to learn language, and they work with this idea until they see evidence in their mother tongue that negates the original assumptions.

I should like, however, to take Bickerton to task on his criticism of Noam Chomsky's view of acquisition. First of all, Bickerton maintains that the acquisition mechanism of transformational grammar is conceived of as "selecting" the appropriate grammar from a range of possible grammars. Actually this notion of selection is a rather outdated one: it was the view of acquisition put forth in Chomsky's *Aspects of the Theory of Syntax* (1965). In current, more sophisticated views of transformational grammar, for example that of Chomsky's *Lectures on Government and Binding* (1981), the notion of selection has given way to one of parameter setting whereby a child, given a priori a set of parameters, or "switches," learns the grammar of his language by setting each parameter appropriately as linguistic data are encountered. This is an important difference: selection is vague from a computational point of view, and there can be many interpretations (some of which would suggest algorithms that are undecidable in finite time) on how this selection is to be made. Parameter setting represents a concrete claim, one that Robert C. Berwick has successfully embedded in an implemented model of acquisition. Thus what Bickerton is criticizing is a view of language learning that is no longer even held.

Second, and more important, Bickerton objects to the Chomskyan approach because in it, he asserts, there is no preferred grammatical model, one the child will aim for and utilize until evidence in the language he is learning forces him to change his model. This assertion is false. The idea that the child has no preferences has been taken as an *idealized* model in transformational grammar.

Similarly, there is the idealization of instantaneity: the obviously factually false model whereby the formal language learner acquires knowledge of his mother tongue by an instantaneous exposure to a large body of data. But these are *not* part of a theory of acquisition. They are (as Chomsky himself puts it) merely idealizations, like that of an ideal gas in physics. In fact, for many years linguists have had a notion of markedness by which specifications on features or parameters could be characterized as being either marked (less likely, or less highly valued) or unmarked (more likely, or more highly valued). One of the central ideas of markedness theory is that the unmarked specification is the preferred one from the point of view of acquisition and the one the child will select unless he has some evidence to the contrary.

Creole languages in this view would represent the maximally unmarked case: having no counterevidence, the child assumes the unmarked settings, which give him a creole grammar. Hence within the transformational-grammar view of acquisition we can explain the effects Bickerton's theory seeks also to explain. Nevertheless, it should be pointed out that the parameter-setting model is quite different from Bickerton's theory, which maintains that the creole grammar is the *only* encoding of linguistic knowledge given to the child. Without a formal statement of exactly what to do given an input of contrary evidence, we are left to guess, in Bickerton's approach, as to exactly how a child does proceed if he is exposed not to Hawaiian Creole but rather to English.

In contrast, Chomsky's theory, given the markedness conventions, makes a specific claim on how this should occur. It does not state that children will not make mistakes, or even start off with creole-like grammars; it does give a formal model (by providing a set of parameters any of which can, if necessary, be set to the marked position) for what exactly a child will do in the common case that his mother tongue is not a creole. Bickerton's approach does not, and thus cannot, account for the obvious ability of children to learn noncreole languages.

In summary, current theories of acquisition by way of parameter setting and markedness are marvelously adapted to accounting for the facts of natural language, creole and noncreole alike.

RICHARD SPROAT

Department of Linguistics
and Philosophy
Massachusetts Institute of Technology
Cambridge

Sirs:

Of the points Richard Sproat raises the first has little substance. I fail to



The Christmas bonus fit for
those who give out the Christmas bonus.

Johnnie Walker®
Black Label Scotch
YEARS 12 OLD

Send a gift of Johnnie Walker® Black Label anywhere in the U.S.A. Call 1-800-528-6148. Void where prohibited.

12 YEAR OLD BLENDED SCOTCH WHISKY, 86 PROOF. BOTTLED IN SCOTLAND. IMPORTED BY SOMERSET IMPORTERS, LTD., N.Y. 1983

Now you can take perfect flash shots with your ears.



The Vivitar 3500 flash is so remarkably easy to use, your only requirement is a good pair of ears.

Our 3500 literally talks you into great photography. First, by telling you with a short "beep" that the flash is charged and ready to shoot. And then, by letting out a longer "beeep" after the shot to tell you the light was correct. (We've even included a light that lets you check everything visually.)

Beep, Beep, Flash **Beeeeeeep!**





The automatic 3500 flash is so simple, about all you do is attach it to your Canon, Minolta, Nikon, Pentax, Olympus, Contax, Yashica or other popular 35mm SLR camera, focus and shoot.

Vivitar®

The result? The 3500 delivers a level of performance that's unheard of. You can take superb automatic shots out to 47 feet.

And bounce light to soften the lighting and eliminate "redeye." And because our compact unit also has zoom capabilities, it will work perfectly with an assortment of lenses.

If it seems incomprehensible that our flash does so much and yet costs so little, we suggest you visit a photo store and try one out.

Hearing is believing.

The 3500 Flash by Vivitar

For a brochure (U.S. only) on our complete family of photographic products, send a self-addressed envelope with 40¢ postage to: Vivitar, 210AD, Box 2100, Santa Monica, CA 90406. Vivitar Corporation 1983. In Canada: Vivitar Canada Ltd./Ltée

see how I misrepresented Chomsky by stating that the child must select an appropriate grammar from the range made available by universal grammar, since the "parameter setting" model of Chomsky's *Lectures on Government and Binding* is, as Sproat admits, simply one of the many ways in which such a selection process might be instantiated. I fully agree with Sproat that this latest model is an advance on the earlier one, which was quite vague, but it is still a selection model. Thus statements to the effect that I am criticizing a view of language learning no longer held are calculated to mislead a nonspecialist audience into supposing there has been some radical change in Chomsky's approach to language acquisition, whereas in fact there has been merely a filling out of an original sketch.

The second point is both more substantive and more interesting, but again Sproat leads off with a misrepresentation. A careful rereading of my article fails to reveal anything that could be construed as a claim that within Chomsky's approach there is no preferred grammatical model. Indeed, I show in my most recent article ("The Language Bioprogram Hypothesis," to appear in *The Behavioral and Brain Sciences*, June, 1984) that on one interpretation creole grammar *could* consist of the unmarked parameter settings in Chomsky's latest model, just as Sproat suggests. The situation is much more problematic, however, than Sproat seems to believe.

In the first place Chomsky himself considers (*Lectures on Government and Binding*, page 9) that his model may not work without "indirect negative evidence," that is, if unmarked rules and structures are not exemplified in the input the child receives, he may select marked alternatives to these. Neither Chomsky nor Sproat, to the best of my knowledge, has carried out the thought experiment of seeing how "indirect negative evidence" would work in the situations I have described. Pidgin input to the first creole generation exemplified few if any rules or structures, marked or unmarked. In such circumstances, according to Chomsky's own statement, the child would be driven to abandon unmarked options and replace them with marked ones. These too, however, would remain unconfirmed. According to this model, then, the child of pidgin-speaking parents could not have learned any language—a conclusion that is patently incorrect.

A more radical argument against the "parameter setting" model is that the hypothesis of a single innate grammar might eventually give us a better theory of syntax and how it is acquired. At present markedness and the various parameter settings of Chomsky's model have no principled foundation; they are not deduced from a central theory but

simply induced from empirical observations of the variety of natural-language data.

If universal grammar consisted of a single syntactic rule system narrowly constrained by a small set of principles, however, Chomsky's parameter settings might fall out quite naturally from relaxing one or another (or varying combinations of) these constraining principles to varying degrees. If this proved to be the case, not only would we have an explanatory theory of parameter settings but also there would be a rational explanation of the acquisition of noncreole languages, something Sproat fails to find in my article.

In fact, even now the position is not quite as Sproat describes it. Both Chomsky's model and mine contain things that are innate and things that have to be learned: in Chomsky's terminology "core grammar" and "the marked periphery." Since all languages have at least some creole characteristics, the child equipped with creole grammar but faced with a noncreole language would by no means be in the helpless position Sproat seems to suppose; he would just be less well endowed and have to do more old-fashioned learning than the Chomsky-equipped child. How much of language is innate and how much is learned is still a very open (and empirical) question, in spite of grandiose talk of "theories... marvelously capable of accounting for the facts."

I do not think, however, that either model is explicit enough at this stage. If I am right in suspecting that the range of natural-language grammars can be derived by relaxing the constraints on creole grammar in various ways, we might get a bit closer to a viable theory of the acquisition of language. The child equipped with creole grammar but doomed by fate to learn language *X* would simply have to observe what constraints were relaxed in *X*, and to what extent, in order to arrive at the core grammar of *X*. If in addition that child was equipped with something like the preemption principles proposed by Steven Pinker ("How Do Children Unlearn Their Mistakes? A Hypothesis from Learnability Theory," presented at the meeting of the Society for Research in Child Development in Detroit in April), which would oblige him to reject the output of the creole grammar wherever it differed from that of the target grammar (instead of simply allowing the two outputs to coexist in his speech), then he would have an apparatus powerful enough to learn any natural language in finite, indeed brief, time.

DEREK BICKERTON

Department of Linguistics
University of Hawaii at Manoa
Honolulu

How to get 336 lenses in your camera bag.



You have two choices. Either you find a bag that's extraordinarily large. Or you find one lens that delivers 336 different focal lengths. Like the 75-410mm zoom lens system from Vivitar.

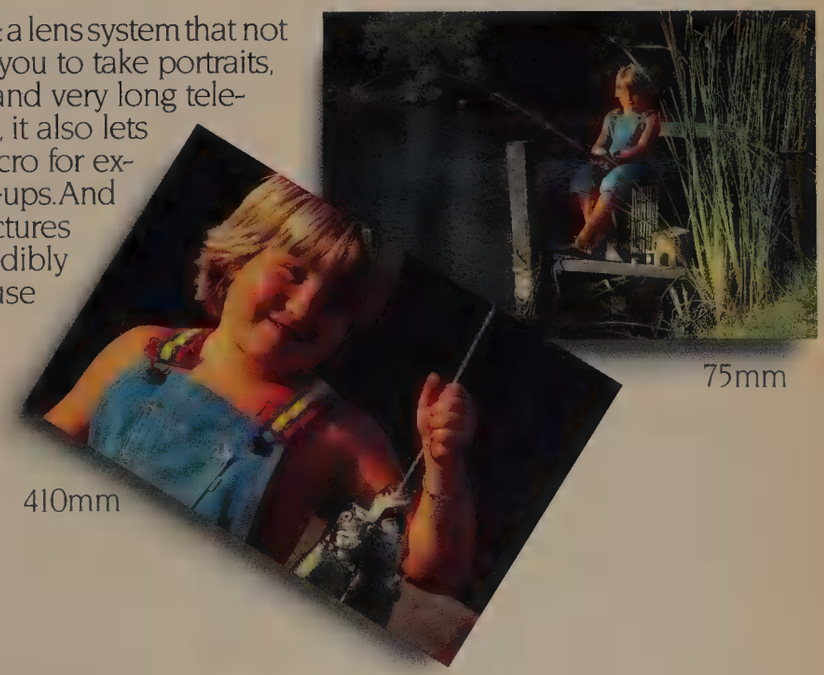
Vivitar®

Our lens system is made up of two components. We start with a 75-205mm f/3.5-4.5 lens, which is the fastest, smallest and lightest lens in its class. And to it, we add a specially designed 2X Matched Multiplier™ for doubling the focal range.

The result: a lens system that not only allows you to take portraits, telephotos and very long telephoto shots, it also lets you use macro for extreme close-ups. And all of the pictures will be incredibly sharp because of the lens' high degree of contrast and resolution.

But performance is not the only part of our system that's unsurpassed. So is its value. There's simply no better 75-410mm lens system at any price. And each system is backed by a 5-year limited warranty.

So the choice is quite simple. You can buy 336 lenses for your Canon, Minolta, Nikon, Olympus, Pentax or other popular 35mm SLR camera. Or you can buy one Vivitar 75-410mm zoom lens system. Ours is certainly a nicer fit in your camera bag.



The 75-410mm Zoom System by Vivitar

For a brochure (U.S. only) on our complete family of photographic products, send a self-addressed envelope with 40¢ postage to: Vivitar, 206AD, Box 2100, Santa Monica, CA 90406. Vivitar Corporation 1983. In Canada: Vivitar Canada Ltd./Ltée.

WHILE SOME CARS MIGHT MAKE IT TO THE NEXT DECADE, STATISTICS SHOW VOLVOS COULD MAKE IT TO THE NEXT CENTURY.

Today people are trying to hold on to their cars longer than ever.

And while statistics show the average life expectancy of today's cars is eleven years, that figure pales in comparison to the life expectancy of today's Volvo.

Because statistics show the average life expectancy of a Volvo is over sixteen years.*

Which could mean over sixteen years of comfortable driving in seats equipped with adjustable lumbar supports that relieve tension and road fatigue.

Years of effortlessly maneuvering through parking lots due to power assisted rack and pinion steering that cuts one of the smallest turning circles of any car.

Years of clean air made possible by a Fresh Air Ventilation System that exhausts stale air from the passenger compartment, and helps prevent odors and fumes from ever getting inside.

If all this sounds like a sensible way to transport yourself to the next century, buy a Volvo.

Who knows?

By the year 2000 we may have a car that'll get you through that century, as well.

VOLVO

A car you can believe in.

*Based upon an actuarial analysis of 1981-1982 U.S. Registration Data conducted by Ken Warwick & Associates, Inc. Due to many factors including maintenance, driving conditions and habits, your Volvo may not last as long. Then again, it may last longer. Summary available at your Volvo dealer. © 1983 Volvo of America Corporation



50 AND 100 YEARS AGO

SCIENTIFIC AMERICAN

DECEMBER, 1933: "The blame for the dearth of satisfactory television transmitters can be laid directly at the door of the Federal Radio Commission. This body has ruled that advertising shall not be a part of any television program and that the latter shall be 'for experimental purposes only.' This kind of caution is commendable in that it may prevent any repetition of the blasts of advertising that have done so much to degrade sound broadcasting. On the other hand, the restriction prevents the operation of self-supporting television transmitters, and without them television receivers become only curiosities, depending for their programs on a few scattered stations operating on perfunctory schedules."

"The doctrine of Nordic superiority is an offshoot of Aryanism, the chief exponent of which was Count Joseph Arthur de Gobineau, a French aristocrat who died in 1882. Gobineau maintained that one race alone, the Aryans, has been the creator and sustainer of all that is great and good in civilization. The idea of an Aryan race was based on the discoveries of similarities in the languages of the Indo-European group, which led to the theory that all these languages were derived from a common stem, the Aryan language. Gobineau and his disciples assumed that the existence of an Aryan language implies existence also of an Aryan race. Having created this mythical race, they attributed to it all virtue and excellence, and they saw in it the source of every great civilization of antiquity and of modern times. The Nordics were represented as descendants of the original Aryans who settled in northern Europe and from whom in turn came the Teutonic and Anglo-Saxon peoples. In spite of all efforts no one has ever been able to produce the slightest bit of real evidence that any such race ever existed. There is no necessary relation between language and race, and the very use of the term 'Aryan' in a racial sense—as the Germans are using it today—has no justification whatever."

"It is very probable that there is a great deal of free hydrogen in Jupiter and Saturn. Hydrogen is enormously abundant in the sun, and these planets have sufficient gravitative power to prevent even so light a gas from diffusing away from their surface into space. The clouds that form the visible surface must arise from the condensation of some substance that comes down at about 100 degrees below zero centigrade. Why these clouds, particularly on Jupiter, show such conspicuous and various colors—white, brown and red—is a problem for the future."

"Drilling an oil well to a depth of more than 11,000 feet, as has recently been done on the edge of California's famous Kettleman Hills, gives rise to three pertinent questions: Will it prove possible to go still deeper? What are the problems of such deep drilling? Will those who risk their fortune find oil in those superheated subterranean rocks? Oil men have been heard to discuss the possibilities of extending their operations down to three miles. They think that they can reach this distant horizon, but that few operators will attempt to go so deep until the present visible supply of oil dwindles to the point where rising prices will justify the high cost."



DECEMBER, 1883: "According to the recent annual report of the Edison Electric Light Company, the First District Works in New York, Pearl Street Station, is now running up to its full capacity. It has 9,811 incandescent lamps in actual use, and it is connected and has wired for 12,379 lamps. Altogether the Edison company now has 246 plants at work and 61,366 lamps."

"The storage and transportation of petroleum is in the hands of two companies, whose pipes cover the entire field of Pennsylvania and convey the oil to reservoirs hundreds of miles distant. The larger of these companies, the United Pipe Lines, is controlled by the Standard Oil Company. Its lines are six in number; two run from Olean to Communipaw, on New York Bay, another runs to Buffalo, one to Cleveland, one to Pittsburg and the sixth to Milton Station on the Reading Road. Its pipes in the aggregate are more than 3,000 miles long, and it owns more than 600 tanks with an aggregate storage capacity of 20,000,000 barrels. The oil that is carried by the pipe lines is crude petroleum. The refining necessary to fit the oil for its commercial uses is done principally at Cleveland, Buffalo, Oil City, Pittsburg and in the vicinity of New York City. The bulk of the petroleum exported is refined oil."

"Two years ago M. Gaston Tissandier, struck by the advantages that are presented by the new dynamo-electric motors from an aerostatic point of view, operated at the Exhibition of Electricity a small directable electric balloon. Since that time he has constructed a new bichromate of potash pile of light weight and capable of yielding considerable electricity for a short time. Soon afterward M. Albert Tissandier began a special study of the mode of constructing a large elongated balloon that should offer the indispensable conditions of lightness and stability in the atmosphere. The resulting balloon is 28 meters in length by 9.2 meters in diameter. The car contains the motor, which is formed of 24 bichromate of potash elements that actuate a Siemens dynamo connected with the propeller through the intermedium of gearing. The motor has a power of 100 kilogram-meters, equivalent to that of 10 men, and drives the propeller, which is about three meters in diameter, at the rate of 180 revolutions per minute. The balloon was inflated on the 8th of October by means of a large hydrogen apparatus. The Messrs. Tissandier ascended slowly at 20 minutes past three o'clock in the afternoon amid the applause of numerous spectators. The balloon remained for a long time over the Bois de Boulogne and was easily swerved to the right and left under the action of the propeller and rudder."

"Sir Charles William Siemens, the well-known scientist, engineer and electrician, died in London on the 20th ult. of rupture of the heart. Dr. Siemens was born at Lenthe, Hanover, in 1823. In 1843 and 1844 he visited England to introduce a gilding and silvering method of his brother Werner's. He ever afterward made England his home, becoming a naturalized citizen in 1859. Between 1856 and 1861 he worked out, with his brother Frederick, the regenerative gas furnace. While working upon this furnace Sir William also sought to make steel and iron direct from the ore. With this object in view he constructed his sample steel works at Birmingham in 1866. Thereafter he continued to produce steel upon the open hearth of his regenerative gas furnace. In 1868 he originated the Landore Siemens steel works, which produce upward of 1,000 tons of cast steel per week. Ever since 1848 Dr. Siemens had been interested in telegraphy and had occupied a prominent position in the development of electrical devices. In 1858 he established, with his brothers Werner and Carl and Dr. Halske of Berlin, the works now known as those of the Siemens Brothers, in London, Berlin and St. Petersburg. He planned and had built the steamer *Faraday* for laying ocean cables, and at their Woolwich factory the brothers manufactured several of the Atlantic cables. In the department of electric lighting he was esteemed second to no one in England. His name is connected with many other inventions."

10 QUESTIONS TO ASK YOURSELF BEFORE YOU BUY A BUSINESS COMPUTER.

The toughest thing about investing in the best computer is finding which computer is best for you. Because computers can be just as different as the people they work for.

That's why we've devised a list of 10 basic, unbiased questions to help simplify the choice of which computer will get along with you.

1 HOW MUCH SHOULD I SPEND?

As little as possible, and still get a fully functioning computer.

That would not include machines you see advertised for under $1,000. They're mainly designed for home entertainment and balancing your checkbook in full color.

For a serious business system, with all the capabilities you need, you can spend as little as $1,600, including software. Or as much as $5,000. And up.

Of course this gives you a choice of more than 500 different makes and models. So rather than devote the next 5 years of your life trying 500 different computers, a good short list would include computers from IBM, Kaypro, Apple and Tandy—the leading brands.

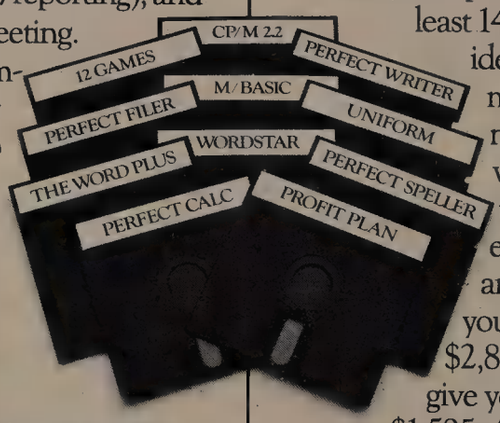
2 WHAT CAPABILITIES DO I NEED?

While businesses can be very different, their computer needs usually aren't.

We've found that 95% of all business needs can be fulfilled by a computer with three basic capabilities. Word Processing/Spelling, Data Base Management (filing/reporting), and Financial Spreadsheeting.

If the same computer has a popular 'operating system', so much the better.

It will have thousands of other programs available to fill more specialized needs.



3 HOW MUCH HARDWARE DO I NEED TO GET STARTED?

Beyond the basic computer, you need a monitor and at least one disk drive, the device that stores information. For word processing or data base management, you will need a computer with 2 drives. Such powerful programs require the storage capacity of their own disk drive.

A second drive also eases copying data from one disk to the other.

4 HOW MUCH MEMORY WILL I NEED?

For most business purposes, a machine with a built-in memory (RAM) of 64K, about 42 typed pages.

Since you'll need a disk drive to store data, its capacity should be at least 140K. Of course, the idea is to get the most memory for the money. For example, when you compare the hardware of equivalent systems, an IBM PC will give you 320K for about $2,800. A Kaypro II will give you 400K for about $1,595. And an Apple IIe will give you 286K for about $2,400. So once again, the choice depends upon your needs.

5 DO I NEED AN 8-BIT OR 16-BIT SYSTEM?

A 16-bit system costs more, gives you a little more speed, and can run longer programs. Unfortunately, only a handful of programs

take advantage of 16-bit capacity.

75% of all microcomputers sold today are of the 8-bit variety, indicative of their ability to satisfy the needs of most businesses.

6 WILL I NEED A PRINTER?

Most people do, sooner or later. It's a fast, efficient way of getting information from your computer screen to other people. For letters, for reports, for balance sheets and for records. That's why it's important to make sure the computer you want has a built-in printer connection. Otherwise when you go to buy a printer, you'll also find yourself paying for an interface.

7 HOW MUCH TRAINING WILL I NEED TO USE MY COMPUTER?

You won't have to learn to write programs to use your computer. Somebody's already done that for you. You will have to learn how your machine operates, which may take a few hours.

And you'll learn how to run programs. The time that it takes depends on their complexity. The ideal way to save time is having a family of programs that share the same commands. Each program in the series will take less time to learn.

8 WHERE CAN I GO FOR GOOD SOLID ADVICE?

Do the same thing you would do if you were shopping for a stereo or an expensive camera. Ask an enthusiast. A computer buff. You probably know one. If not, call a consultant. In the long run, he or she could save you a lot of money by directing you to the right combination of hardware and software. And by advising you whether to buy a computer now or wait for prices to come down further.

It is true that the price of serious business computers is drifting down an average of 10-12% a year. The average system costing $3,000 today may cost $2,700 next year. Yet, you could lose thousands of dollars of increased productivity, waiting to save $300.

SPECIFICATIONS

Microprocessor
Z-80
Operating System
CP/M 2.2
User Memory
64K
Disk Drives:
2 drives, 400K, unformatted
Interfaces
1 Serial
1 Parallel
Keyboard
Detached, 63-key with numeric keypad
Software included:
Perfect Writer word processing
Perfect Speller
Perfect Filer
Perfect Calc spreadsheet
Wordstar word processing
The Word Plus
Profit Plan spreadsheet
M-Basic
12 Games
Uniform—allows computer to 'read' and 'write' TRS-80, Osborne, Xerox disks
Dimensions
Height: 8 inches
Width: 18 inches
Depth: 15½ inches
Weight: 26 lbs. (portable)

9 WHY IS THERE SUCH A GREAT DIFFERENCE IN THE PRICE OF COMPUTERS?

Some differences are based on capabilities. Business computers capable of playing joystick games cost more than ones that stick to business. And, as we mentioned before, 16-bit computers cost more than 8-bit systems.

A more important difference is the way computers are manufactured and sold. Most start with the basic keyboard and microprocessor, for a basic price.

You then pay more for a monitor. More for disk drives. More for communication interfaces. More for software. All optional extras that can run up the price two or three times.

However, at Kaypro we don't consider monitors, disk drives and software to be optional extras. So we make and sell completely integrated systems with all the hardware you need. All the software you need. All for $1,595.

Of course, a lot of other computer companies feel you should pay extra for everything you get.

10 WHICH IS THE SMARTEST COMPUTER FOR ME TO BUY?

All things considered, we'd consider it a Kaypro. And since this is a Kaypro ad, you wouldn't think we'd suggest anything else. But fortunately for us, our computer gives us a lot to be biased about. Off the shelf, it gives you everything you want for 95% of your business needs. Plus the ability to run 3,000 more programs. It's the complete business computer.

What do you get complete? 64K RAM, Z-80 microprocessor. CP/M, the most widely used business software operating systems. A 9" easy-to-read, green monitor. Two disk drives (400K capacity). Interfaces for both a printer and communications. A detachable keyboard, with a numeric keypad. And one of the leading computers in the $1,000 to $5,000 price range.

A Kaypro may not be the best computer for everyone. But for $1,595 complete, it gives you the best solution to serious business problems.

CALL 800-447-4700 FOR THE DEALER NEAREST YOU.

Registered Trademarks: Apple — Apple Computer, Inc., IBM — IBM Corp., CP/M — Digital Research, Inc., Z-80 — Zilog, M-Basic — Microsoft, Inc., Tandy, TRS-80 — Tandy Corporation, Osborne — Osborne Computer Corporation, Xerox — Xerox Corporation. Prices based on published information as of July 15, 1983. © 1983 Kaypro Corporation.



KAYPRO

The Complete Business Computer.

THE AUTHORS

JOHN D. HARBRON ("Modern Icebreakers") is a foreign affairs analyst for the Thomson Newspaper Group in Canada. His career as a professional writer has been closely connected with his long-standing interest in ships and the sea. After being graduated from the University of Toronto and doing postgraduate work at the University of Havana he entered the Royal Canadian Navy. His first navy job was as naval officer and professor of naval history at the Canadian Naval College at Royal Roads, B.C. After retiring from the service with the rank of commander he worked for *Business Week* and the *Toronto Telegram* and served as correspondent for several British and Canadian shipping publications; he left the *Telegram* to join the Thomson group. Harbron is a recipient of the Maria Moors Cabot Medal of the Columbia University Graduate School of Journalism and a commander of the Order of Isabel la Católica of Spain.

TORRENCE V. JOHNSON and LAURENCE A. SODERBLOM ("Io") are respectively an astronomer and a geophysicist; they were members of the National Aeronautics and Space Administration teams that obtained the first closeup views of many bodies in the solar system. Johnson got a B.S. at Washington University and a Ph.D. in planetary science from the California Institute of Technology. He was on the staff of NASA's Planetary Astronomy Laboratory and the National Research Council before taking up his present job as senior research scientist at the Jet Propulsion Laboratory of Cal Tech. He was a member of the outer-planets-probe working group and the Voyager imaging-science team. Soderblom received his Ph.D. in planetary science and geophysics at Cal Tech; he received his undergraduate education at the New Mexico Institute of Mining and Technology. Since getting his doctorate he has worked for the U.S. Geological Survey, in which he holds the rank of supervisory physical scientist.

PETER S. WELLS ("An Early Iron Age Farm Community in Central Europe") is associate professor of anthropology at Harvard University and associate curator of European archaeology at the Peabody Museum. He has been associated with Harvard almost continuously since his undergraduate days. He got a B.A. at Harvard College in 1970 and went on to earn his Ph.D. in anthropology from Harvard University in 1976; his graduate studies included work at the University of Tübingen. His main research interest is economic and social change in the Bronze Age and the Iron Age. The excavations in Bavaria that are described in his article are currently being expanded to include other prehistoric settlements in the region to yield a broader picture of the development of society in central Europe in the final millennium B.C. Wells's research is supported by the National Geographic Society, the National Science Foundation, Earthwatch and the Center for Field Research.

RICHARD E. DICKERSON ("The DNA Helix and How It Is Read") is professor of biochemistry and geophysics and director of the Molecular Biology Institute at the University of California at Los Angeles. He attended the Carnegie Institute of Technology as an undergraduate before receiving his Ph.D. in physical chemistry from the University of Minnesota. After a year of postgraduate study at the Cavendish Laboratory of the University of Cambridge he returned to the U.S. to take up an appointment at the University of Illinois. After four years there he moved to Cal Tech. In 18 years at Cal Tech his main research interests were analyzing the crystal structure of the electron-transport protein cytochrome *c* and the study of the evolution of molecules. Since going to U.C.L.A. in 1981 his interests have shifted to the analysis of the structure of short stretches of DNA and the complexes such molecular structures form with antitumor drugs and the proteins that act to repress the reading of the double helix.

BRYCE S. DEWITT ("Quantum Gravity") is professor of physics at the University of Texas at Austin and director of the Center for Relativity there. He holds three degrees from Harvard: a B.S., an M.A. and a Ph.D. in physics. After working briefly at several institutions he became senior research physicist at the Radiation Laboratory of the University of California at Berkeley. He left Berkeley to join the department of physics at the University of North Carolina at Chapel Hill. He was at North Carolina for 15 years, leaving in 1972 to move to Texas. In addition to these positions he has held visiting appointments at many institutions, including the École d'Été de Physique Théorique at Haute Savoie in France, the University of Osaka, Stanford University, the University of Oxford and the Institute for Theoretical Physics in Santa Barbara.

ARTHUR T. BERGERUD ("Prey Switching in a Simple Ecosystem") is research associate at the University of Victoria in British Columbia. He got a B.S. at Oregon State College in 1956 and an M.S. in biology at the University of Wisconsin five years later. After receiving his master's degree he went to Newfoundland to join the provincial Department of Wildlife. In his 11 years with the department he rose from district biologist to director of wildlife management. In 1966 he moved to the University of British Columbia to continue his formal education. He earned his Ph.D. in biology from British Columbia in 1968; his thesis was on the caribou population of Newfoundland.

ALBERT C. GROSS, CHESTER R. KYLE and DOUGLAS J. MALEWICKI ("The Aerodynamics of Human-powered Land Vehicles") share an interest in the topic of their article. It is one they approach from quite different professional perspectives. Gross, a management consultant and journalist, is president of Action Research Associates, Inc. He has two bachelor's degrees: one in American history from Columbia College (1969) and one in psychology from San Diego State University (1975). After getting his second degree he continued the study of experimental psychology at the University of California at San Diego, which granted him an M.A. in the subject in 1976. Kyle is professor of mechanical engineering at California State University at Long Beach. His bachelor's degree is from the University of Arizona; his M.S. and Ph.D. in engineering are from the University of California at Los Angeles. Outside the laboratory his avocation is the aerodynamics of human-powered vehicles, a subject on which he has often published. He is cofounder and past president of the International Human Powered Vehicle Association (IHPVA). Streamlined bicycles of his design have been employed in setting five world speed records for human-powered vehicles. Malewicki is director of Aerovisions, Inc. He has an M.S. in aeronautical and astronautical engineering from Stanford University. He is a consultant who works mainly for Aerochem, Inc., on aerospace projects, including ones related to the space shuttle. His work has included stress analysis of two-man submarines and performance analysis of hot-air balloons designed to set altitude records. He has also designed motorcycles with which distance and altitude records for jumping have been set.

DONALD D. HOFFMAN ("The Interpretation of Visual Illusions") is assistant professor of cognitive science and computer science at the University of California at Irvine. He got a B.A. in 1978 from the University of California at Los Angeles before going on to earn his Ph.D. in computational psychology in 1983 from the Massachusetts Institute of Technology.

COMPUTER RECREATIONS

On the finite-state machine, a minimal model of mousetraps, ribosomes and the human soul

by Brian Hayes

The most powerful computers have neither hardware nor software; they are built out of pure thought stuff. Among these abstract machines the most celebrated is the one invented in 1936 by the British mathematician Alan Mathison Turing. It can do more than any computer made of mere silicon ever could; indeed, it can compute anything that can be computed. A related class of conceptual computers lack the omnipotence of the Turing machine, but they are no less interesting. They are called finite-state machines or finite-state automata, and they establish the minimum specifications of a working digital computer.

Properly defining a finite-state machine calls for a degree of mathematical rigor that is not appropriate here. The nature of the concept can be made clear, however, by means of a few examples. When I went out looking for finite-state machines, I found an excellent specimen in a station of the Lexington Avenue subway in New York. It is a turnstile, an old one made not with the compact steel tripod of current practice but with four oak crossarms, worn smooth by a river of hands and hips.

The turnstile has two states: locked and unlocked. Suppose it is in the locked state, so that the arms cannot be turned. Putting a token into the slot alters the internal mechanism in some way that allows the arms to move; in other words, the token induces a transition to the unlocked state. Rotating the arms by 90 degrees causes another transition that restores the turnstile to the locked state. The transitions are shown schematically in the upper illustration on the next page. The states of the system are represented by nodes (boxes) and the transitions by arcs (arrows) between them.

In the finite-state analysis of the turnstile, inserting a token and pushing on the arms are the possible inputs to the system. The response of the machine depends both on the input and on the state at the time of the input. Pushing on the crossarm when the turnstile has not yet received a token will not get you a ride on the subway. Inserting a token when the arms are already unlocked is also futile, although in a slightly different way. The second token is accepted, but it has no effect on the state of the machine; one person is admitted and then the turnstile locks again. Three or four tokens in sequence are likewise accepted but buy only one ride. Skeptics may want further evidence before accepting the generalization that all tokens after the first have no effect, but they will have to supply their own tokens.

The reason the turnstile cannot give multiple rides for multiple tokens is that it has no means of counting the tokens it has received. Its only form of memory is a rudimentary one: by changing from one state to the other it "remembers" whether the most recent input was a token or a push on the crossarms. All earlier inputs are lost. It is worth noting that this forgetfulness can never work to the disadvantage of the city. It could be worse: a turnstile could be designed to change state after every token, regardless of the present state, in which case two tokens in a row would admit no one.

The turnstile illustrates most of the essential properties of a finite-state machine. Obviously the machine must have some states, and there can be only a finite number of them. There can be inputs and outputs associated with any state. The states must be discrete, or clearly distinguishable, and the transitions between them must be effectively instantaneous. In these matters much depends on the point of view: day and night are discrete states if one is willing to define sunrise and sunset as instantaneous processes. The set of states, the inputs and the outputs constitute the entire machine; there can be no auxiliary devices, and in particular no facilities for the storage of information.

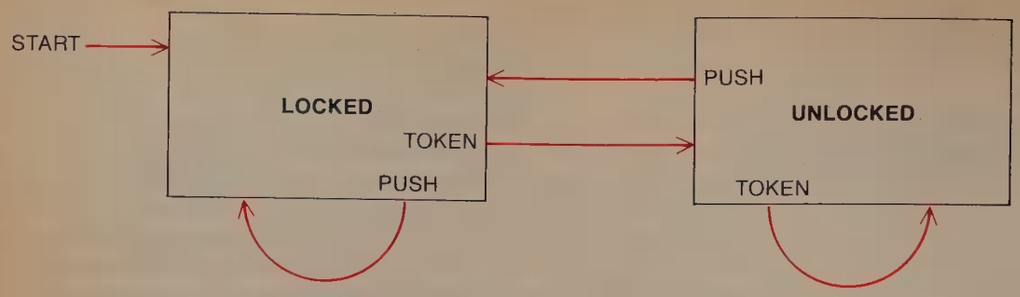
The rules for building a finite-state machine allow some scope for variation. There are deterministic and nondeterministic machines, Moore machines and Mealy machines. In a deterministic machine a given input in a given state invariably has the same result; in a nondeterministic machine there can be several possible transitions. In the Moore machine (named for Edward F. Moore) each state has a unique output. In the Mealy machine (named for G. H. Mealy) the outputs are associated with the transitions rather than the states. It turns out, however, that the variety of architectures is something of an illusion. Any task that can be done by one kind of finite-state machine can be done by the other kinds as well, although the number of states needed may vary. Here I shall discuss mainly deterministic Moore machines, which have the simplest structure.

When you start looking for finite-state machines, you find them everywhere. Coin-operated devices are favorite textbook examples. Some vending machines are less rapacious than the subway turnstile: once they have received the proper amount of money they enter a state in which all additional coins are rejected. The coin-operated device with the largest number of possible states is surely the Las Vegas slot machine. In principle it is deterministic, but finding an input (a coin and a pull on the handle) that will cause a transition to a particular final state is nonetheless challenging.

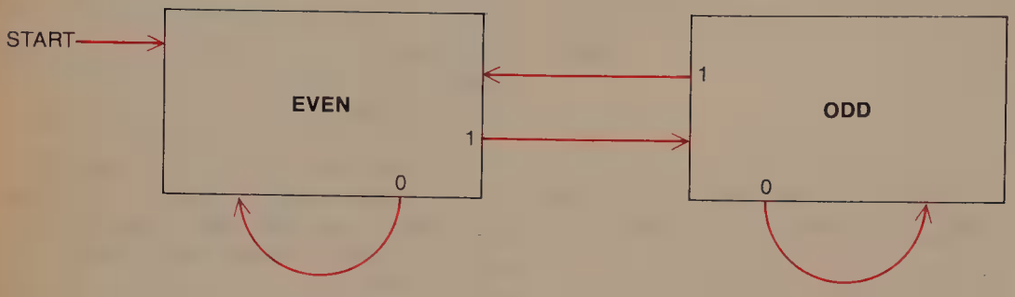
Many household appliances can be regarded as finite-state machines, although they tend to be rather dull ones. A clothes washer goes through an inflexible sequence of states—filling, agitating, rinsing, spinning—and the few meaningful inputs, such as pulling the plug out of the electric outlet, generally have the same effect in all the states. Similarly, a traffic light has a small repertory of states, which repeat indefinitely. To me the most boring of all finite-state machines is a digital clock. If it displays the month, the date and the passage of hours, minutes and seconds, it has some 31 million states; in the course of a year it visits each state exactly once.

A mousetrap is a finite-state machine; the mouse, usually to its misfortune, triggers a transition from the cocked state to the sprung state. A combination lock is a finite-state machine with many possible inputs, only one of which causes a state transition. A telephone has states that might be labeled on hook, off hook, waiting, dial tone, dialing, ringing, connected and out of order. An automobile can demonstrate vividly that the effect of an input varies according to the present state of the system. What happens when you press the accelerator pedal to the floor? It depends. Is the engine running? Is the clutch engaged? Is the parking brake off? Is the transmission in gear? Is it in forward or reverse? Is the garage door open?

In the living cell the molecular system made up of the ribosome and the vari-



A state-transition diagram for a subway turnstile



The parity-testing machine

ous species of transfer RNA operates as a finite-state machine. The inputs are the four nucleotide bases of messenger RNA, designated by the abbreviations *U, A, G* and *C*. The outputs are the 20 amino acid components of proteins. A chain of nucleotides is recognized as a valid input to the machine only if it begins with the "start" signal *AUG*. Thereafter the machine reads the input stream continuously, changing state as each codon, or triplet of nucleotides, is recognized. The three special codons *UAA, UAG* and *UGA* are "stop" signals: when one of them is encountered, the machine halts. Many other biological systems can usefully be represented as finite-state machines; examples that come to mind are the hemoglobin molecule and the promoter and repressor proteins of bacteria.

In the theology of Thomas Aquinas the soul is a finite-state machine, a wonderfully elaborate and fully deterministic one. It is created in a state of jeopardy, as a consequence of original sin. On baptism it enters a state of grace, but certain acts (idolatry, blasphemy, adultery and so forth) induce a transition to a state of sin. Confession, repentance and absolution are then needed to restore the soul to grace. The effect of a final input, death, depends critically on the state of the soul at the moment of death: in a state of grace death leads to salvation but in a state of sin it leads to damnation. The soul machine is actually more complicated than this description suggests. A full account would have to distinguish among the various grades of sin (venial and mortal, actual and habitual) and would have to include other possible states of the soul (such as those associated with limbo and purgatory) and other possible inputs (such as the Last Judgment).

In quantum mechanics even the atom becomes a finite-state machine, and hence so does everything that is made up of atoms. The states of the atom are the allowed energy levels; the inputs and outputs are photons, or quanta of electromagnetic radiation. In a precise description I think the atom would be classified as a nondeterministic Mealy machine with epsilon transitions. It is nondeterministic because the effect of an input cannot be predicted with certainty. It is a Mealy machine because the nature of the output (namely the energy of the photon) is determined by the transition, not by the state entered. Epsilon transitions are those that can take place in the absence of any input; they must be included in the model because an atom can emit a photon and change its state spontaneously.

Is the brain a finite-state machine? As it happens, the modern study of finite-state systems began with a model of neural networks introduced in 1943 by Warren S. McCulloch and Walter Pitts. The neurons of McCulloch and Pitts were simple cells with excitatory and inhibitory inputs; each cell had a single output and two internal states: firing and not firing. The cells could be arranged in networks to carry out various logic functions, including the "and," "or" and "not" functions that are now commonplace elements of electronic logic systems. The equivalence of the idealized neural networks to state-transition diagrams of the kind shown here was established in 1956 by Stephen C. Kleene of the University of Wisconsin at Madison.

Forty years after the work of McCulloch and Pitts it is still subject to dispute whether the brain can reasonably be classified as a finite-state system. Of course the number of neurons is necessarily finite, but that is not the only issue. A real neuron is far more complicated than a two-state cell, and some of its properties may vary over a continuous range rather than being constrained to occupy discrete states. Furthermore, the prohibition of auxiliary information storage in a finite-state model of the brain is awkward at best. If mental life is no more than a succession of instantaneous states, without knowledge of its own history, then what is memory?

The states of mind discussed in psychology, such as boredom, fear, thirst, ecstasy and grief, seem to fit more readily into the apparatus of a finite-state theory. On the other hand, the states are so numerous and the transitions are so poorly understood that the model is useless. Only for lower animals is it possible to draw more than isolated fragments of the state-transition diagram, and in those species the experimenter can have no direct access to the presumed mental states. Indeed, much work of this kind has been done by behaviorists who deny the very existence of mental states.

The case of the digital computer—and here I mean the tangible machine, the hardware—is also problematic. The common mental model of a computer, formulated by John von Neumann, divides the machine into a central processing unit and an array of memory cells. There is no doubt that the finite-state concept can be applied to the various components of the central processor, such as registers, adders and the control mechanism that directs the internal operations of the processor.

The trouble begins when the memory is taken into account. Under the rules for building a finite-state machine no external memory is allowed, and so each cell must be viewed not as a storage facility separate from the processor but as a part of the overall machine state. If all the cells are blank, the computer is in one state; if a single cell is filled, another state is entered, and so on. This conception of the computer is singularly unilluminating, in part because it makes no connection between the state of the machine and what it is doing. Moreover, the number of states is immense. Even a computer of trivial size (100 binary elements), running continuously throughout the age of the universe, could not possibly have worked through all its states.

The primary role of the finite-state machine in computer science is at a higher level of abstraction than the clockwork mechanisms of the hardware. A computer running under the direction of a program is no longer an assemblage of logic gates, registers, memory cells and other electronic paraphernalia; it is a "virtual" machine whose working parts are defined by the program and can be redefined as neces-

sary. Whereas the hardware knows only binary integers and simple commands for moving and manipulating them, the virtual computer deals with far more expressive symbol systems: words, equations, arrays, functions, vectors, codons, lists, images, perhaps even ideas. Finite-state techniques can be valuable in creating the virtual computer, and sometimes the virtual computer *is* a finite-state machine.

Consider a program whose object is to read a series of binary digits (1's and 0's) and report whether the number of 1's received is even or odd. (The task has practical significance; for example, such parity-checking programs are employed to detect errors when digital data are transmitted by telephone.) The program can be constructed as a finite-state machine with two states, as is shown in the lower illustration on the opposite page. Operation begins in the even state, because initially no 1's have been received and 0 is considered an even number. Each 1 in the input stream causes a change of state, whereas a 0 received in either state leaves the state unchanged. Even though the machine cannot "remember" any inputs before the most recent one and certainly cannot count the 1's or 0's, its output always reflects the parity of the input stream.

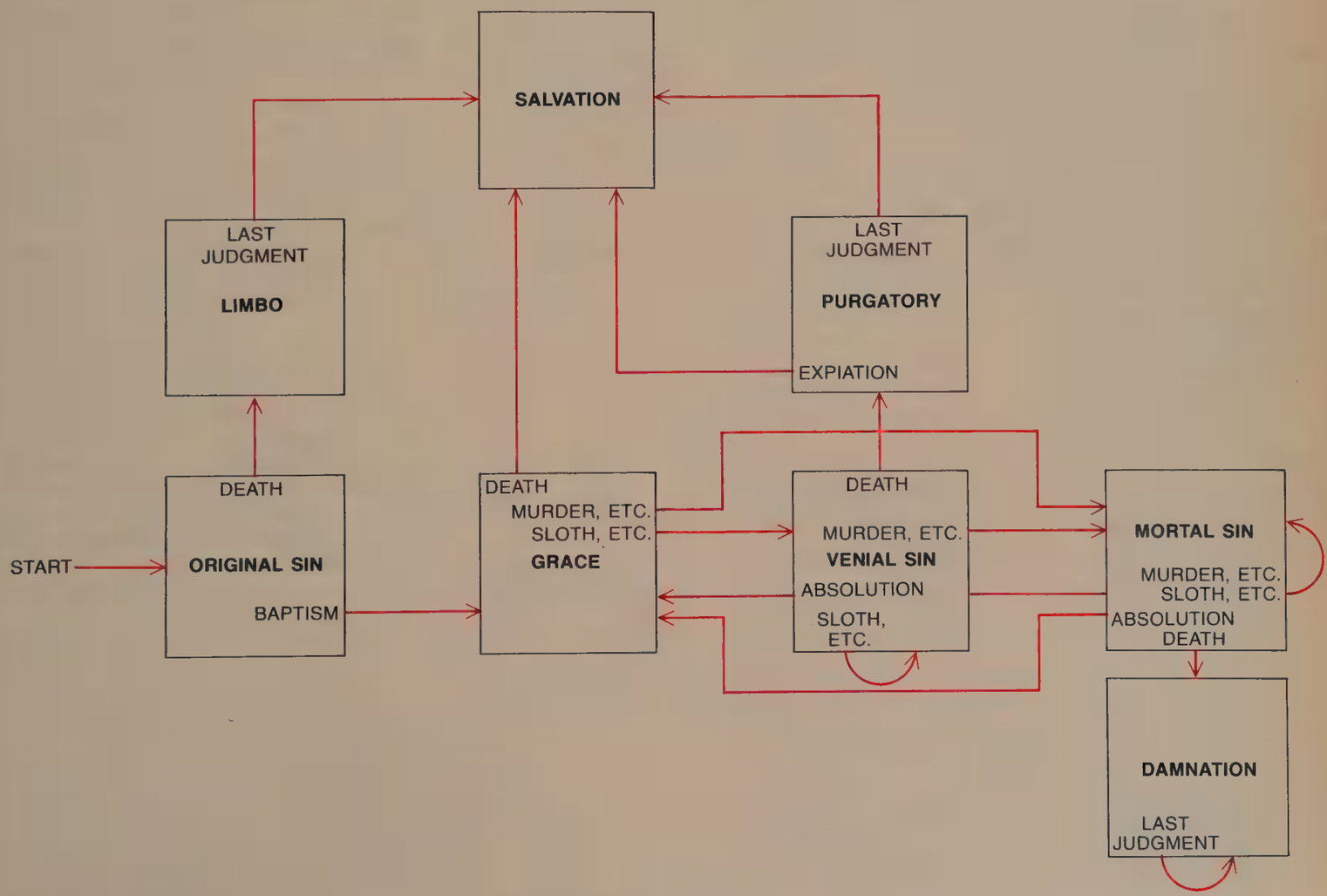
The finite-state model of computation is commonest in programs that deal in some way with text or other information that takes a linguistic form. The preeminent example is found in compilers: programs that translate programming statements in a source language into equivalent statements in a target language, most often the "machine language" of a particular computer. Compilers and other translating programs are essential to the notion of the virtual machine; they mediate between symbols with human meaning and those recognized by the computer.

The part of a compiler that can be designed as a finite-state machine is called the lexical scanner. Like the subway turnstile, it is a token-gobbling device. In this case, however, the tokens are the words, or fundamental lexical units, of the language. The scanner examines each group of characters and determines whether it is a genuine token, such as a command or a number; if it is not, the scanner rejects it as nonsense, just as the turnstile would reject a slug.

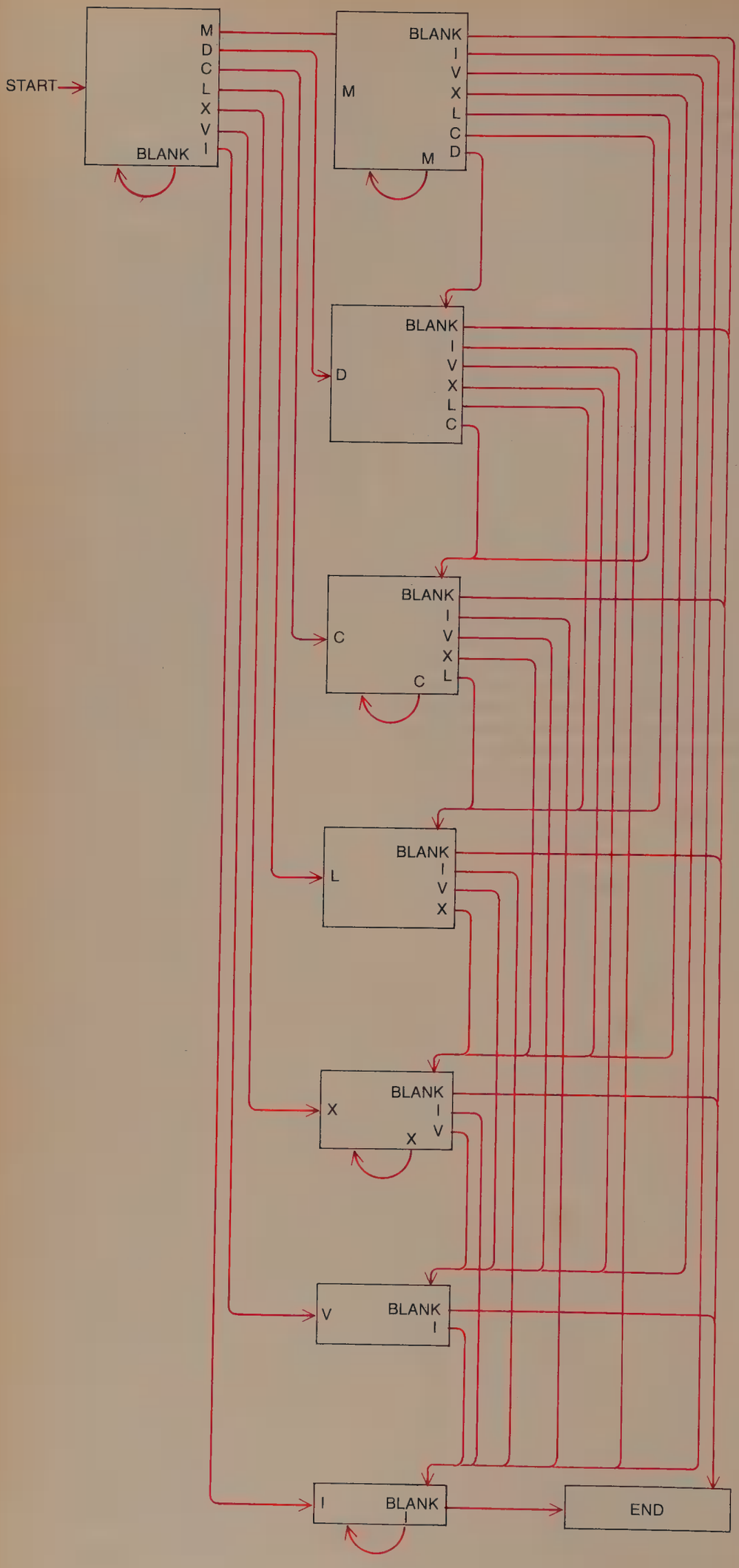
The operation of a lexical scanner can be illustrated by a finite-state machine designed to recognize the tokens of a simple language, albeit one of limited expressive range: the tokens consist exclusively of Roman numerals. Indeed, only Roman numerals of a special form are accepted; they must be given in strict additive notation, so that 9 is represented by VIIII rather than by IX. (There is evidence that the Romans themselves employed the additive notation; the subtractive form is thought to have been a German innovation.)

A state-transition diagram for the Roman-numeral machine is shown in the illustration on the next page. Its alphabet of input symbols includes the letters M, D, C, L, X, V and I as well as the space symbol, or blank. Any initial blanks are simply ignored, but once the first letter is received the program makes an immediate transition to a state identified (for convenience) by the name of the letter. If the first letter is an M, it can be followed by any character from the allowed set, including another M. If the next character is a D, however, the situation is different. From the D state no transition back to the M state is defined, because any series of symbols that includes DM cannot be a well-formed token in the language of additive Roman numerals. Furthermore, there is no transition from the D state to the D state itself, so that DD is also an excluded sequence. (The reason is that the "half value" symbols D, L and V cannot be repeated in proper Roman numerals.)

In the D state the only recognized letters are the lower-valued ones C, L, X, V and I. The same set is accepted in the C



States of the soul in the theology of Thomas Aquinas



A lexical scanner for a language of Roman numerals

state (because C can be repeated), but in the L state only the letters X, V and I are recognized. The rule governing the transitions should be clear. The states are arranged in a hierarchy, and once a given level has been reached the machine can never return to a higher level; in the half-value levels it cannot even remain at the same level. By the time the I state is reached only an additional I or a blank is allowed. The blank, entered at this point or at any other time after the first letter, indicates the end of the token and sends the machine back to its starting state, ready to receive the next Roman numeral.

No programming language known to me allows numbers to be entered in Roman form, but virtually all such languages have facilities for handling Arabic numbers. The techniques for recognition are similar, although there is a greater variety of formats. Simple integers such as 137 can be handled in principle by a one-state machine, but the several parts of a number such as $+6.625 \times 10^{-27}$ require a more elaborate lexical analysis.

The ribosome–transfer-RNA system can be regarded as a lexical scanner that recognizes biologically meaningful nucleotide sequences in a molecule of messenger RNA. To be accepted a sequence must begin with a start codon and end with one of the three stop codons; between these boundaries any combination of the input symbols *U, A, G* and *C,* taken three at a time, is allowed.

Lexical analysis is only the first step in the process of compilation. The components of the compiler that are called into action after the lexical scanner are the parser and the code generator. The parser takes as its input the tokens identified by the scanner and analyzes their syntactic relations; this is the closest the compiler comes to understanding the meaning of the program statements it translates. The code generator writes a program in the target language that carries out the functions specified by the parsed statements.

For the toy languages considered here the tasks of the parser and the code generator are trivial. The compiled form of a statement in the Roman-numeral language might be simply the Arabic equivalent of the number. It could be generated by the following strategy. Before a token is scanned a storage cell is specified and is set equal to zero. Then each time the scanner enters the M state 1,000 is added to the value in the cell; for the D state 500 is added, and so on. When the scanning is complete, the memory cell holds the value of the Roman numeral. Note that the toy compiler is no longer a pure finite-state machine, because it has auxiliary storage.

A compiler for the genetic code is even simpler and can be realized entire-



RARE

BLENDED SCOTCH WHISKY

100% SCOTCH WHISKIES

BLENDED AND BOTTLED IN SCOTLAND BY

JUSTERINI & BROOKS LTD

St. James's Street, London, England

WINE MERCHANTS TO THEIR LATE MAJESTIES

KING GEORGE III

KING GEORGE IV

KING WILLIAM IV

QUEEN VICTORIA

KING EDWARD VII

KING GEORGE V

KING GEORGE VI

AND TO HIS LATE ROYAL HIGHNESS

THE PRINCE OF WALES (1921-1936)

J&B

To have J&B sent as a gift anywhere in the U.S., call this toll-free number: **1-800-528-6148.** In Arizona call: **957-4923.**

86 Proof Blended Scotch Whisky. © 1983 The Paddington Corp., NY

Hello, information.

Or, how the IBM Personal Computer can bring you the world.



Modem shown not supplied by IBM.

There's a world of information just waiting for you. But to use it, study it, enjoy it and profit from it, you first have to get at it.

Yet the facts can literally be right at your fingertips—with your own telephone, a modem and the IBM Personal Computer.

Plug into the network.

The modem that plugs into your phone takes the codes your IBM Personal Computer understands and turns them into signals that can be transmitted over the phone lines.

On the other end of the phone lines are independent information services* ready for you to access from your home, school, lab or office. So, in minutes, you can tap the resources of over a thousand data bases—and plug into a world of possibilities.



Something for everyone.

At home, for example, you can shop through an electronic catalog. Access mortgage amortization and tax depreciation schedules. Study the wines of France, the foods of Italy. Find property to sell, rent, swap or share. And play hundreds of games.

At school, you can bring news into the classroom—as it happens. You can teach and learn with facts about everything from phrase origins to plane geometry. You can also tap instructional courses and learn how to program the very IBM Personal Computer you're using.

In the lab, you can call up a wealth of mathematical and statistical programs or retrieve scientific and technological data.

In the office, you can plug into the commodities market. Read abstracts of leading publications. Even get census figures to see who's in your major market. And when you're planning a business trip, your IBM Personal Computer can provide airline schedules and car rental information, as well as tell you what the weather is and where the best restaurants are.

Calling all computers.

And these outside data bases aren't your only information link, either.

With assigned passwords, you can also have access to data that's stored in your company's IBM minicomputer or mainframe via your IBM Personal Computer.

You can even exchange a sales forecast with a manager two floors down or two thousand miles away—one IBM Personal Computer to another.

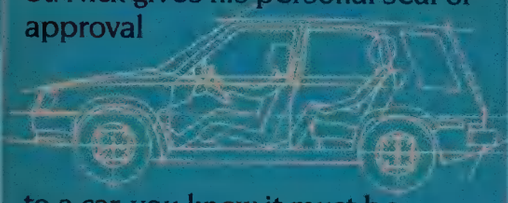
So the IBM Personal Computer really *can* let you call the world. But there's one call to make first. For more information on where to buy the IBM Personal Computer, call 800-447-4700. In Alaska or Hawaii, 800-447-0890. IBM®

The IBM Personal Computer
A tool for modern times

*formation services are available by subscription at additional cost, and are not supplied by IBM.

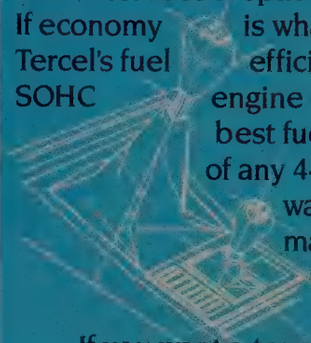
© 1983 Toyota Motor Sales, U.S.A., Inc.

When a man as discerning as St. Nick gives his personal seal of approval



to a car, you know it must be something to say ho ho ho about. Santa's a man with tough requirements. That's why he chose the Tercel 4WD Wagon. The sleigh is fine for Christmas Eve, but for the rest of the year he needs something more practical. He needs a car roomy enough to carry feed for all his reindeer, yet economical and dependable to pull through treacherous roadways.

If the Tercel 4WD Wagon can fit needs as demanding as Santa's, just think how easily it can accommodate you. If room is what you want, stretch out. No small 4WD wagon can match its roominess. The roof line is higher than any other small wagon, for more headroom. And there's a panoramic rear window for more visibility. Five people can sit comfortably with lots of space for gear in the carpeted cargo area. And if you want more luggage space, just fold down the rear seat backs and get 59.8 cubic feet of space.



If economy is what you want, Tercel's fuel efficient 1.5 liter SOHC engine delivers the best fuel economy of any 4-wheel drive wagon: 41 Estimated Highway, (31) EPA Estimated MPG*.

If you want a tough, dependable car, look no further. The Tercel 4WD Wagon can pull through almost any road condition from the North Pole to Main St.** And it switches from front-wheel drive to four-wheel drive on command.

OH WHAT A FEELING! TOYOTA

New for 1984 is a conventional 2WD Wagon that has all the features Santa wants in his 4WD, but it's for people who don't have to go to the North Pole.



Is the Tercel 4WD Wagon for you? Just ask Santa.

*Remember: Compare this estimate to the "EPA Estimated MPG" of other cars with manual transmission. You may get different mileage, depending on how fast you drive, weather conditions and trip length. Actual highway mileage will probably be less than the "Highway Estimate."
**Optional chassis gravel shields recommended for rough road use.

BUCKLE UP—IT'S A GOOD FEELING!

THE 1984 TERCEL 4WD WAGON. THE OFFICIAL CAR OF SANTA CLAUS!



HO HO HO!

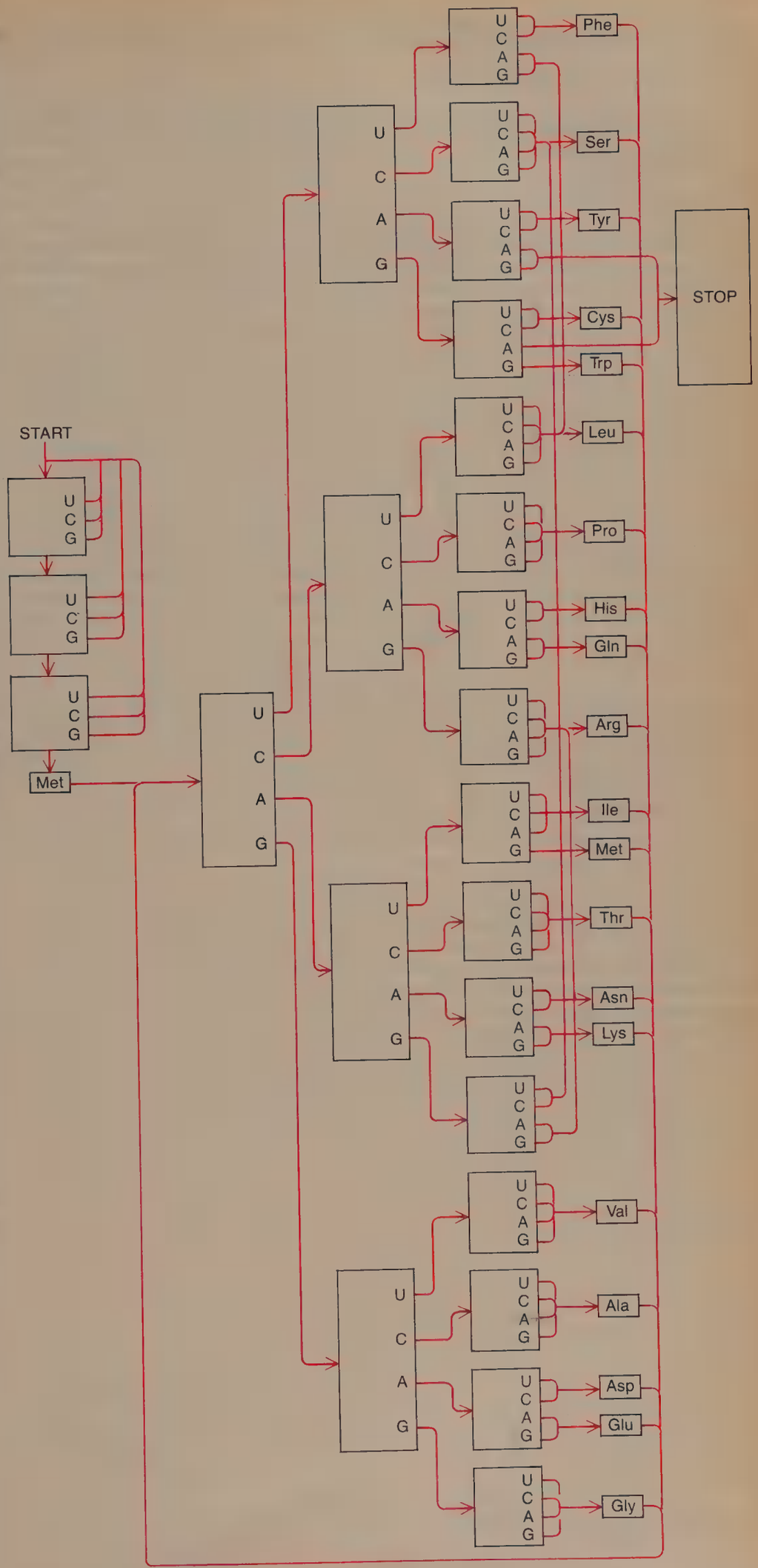
ly within the context of a finite-state system. The compiled "program" is a sequence of the standard three-letter symbols for amino acids; the symbols can be generated as the output of the states of the scanner that recognize codons. The three states corresponding to stop codons have no output.

Creating a compiler for a language large enough to be of general utility is not a casual undertaking, but the underlying architecture of the finite-state machine can at least provide an organizing principle. If the syntax of the language is specified with sufficient precision, part of the work can even be mechanized: it can be done by a compiler compiler, a program whose input is a formal description of a language and whose output is another program that translates statements in the language. As far as I know no one has yet written a compiler compiler compiler.

The identification of tokens by a lexical scanner is in itself a kind of parsing, and the set of all possible sequences of symbols in a token is a kind of language. Indeed, it is an infinite language: unless some artificial limit is put on the length of individual sequences, an infinite variety of recognizable tokens can be formed. How can a machine with only a finite number of parts recognize an infinity of well-formed statements and exclude an infinity of ill-formed ones? The key is in the structure of the language itself. If the statements of an infinite language are to be recognized by a finite-state machine, they must be formed according to strict rules.

The rules were set forth by Kleene in 1956; they define a class of languages called regular languages or regular sets. Kleene proved that a finite-state machine can recognize a language only if it is regular, and further that every regular language can be recognized by some finite-state machine. What is meant by regular can be indicated briefly (although not rigorously) by two rules. First, any finite language is regular and therefore can be recognized by a finite-state machine; after all, one could build a machine with a state for each possible expression of the language. Second, if a language is infinite, it must be possible to parse all its statements by reading one symbol at a time from left to right, or beginning to end, without backtracking or looking ahead. If the acceptability of any symbol is contingent on the presence of another symbol, the governing symbol must be the one immediately to the left.

The second rule is a direct consequence of the limitations of a finite-state machine, which can neither foresee its future states nor keep a record of its past ones; it must choose a state transition based only on the current state and the current input symbol. It is for this



A finite-state machine translates the genetic code into protein

reason that the subtractive notation for Roman numerals cannot be handled by a finite-state machine. If the expression XI is read and the machine interprets it as 11, it cannot go back to revise the value when the next character turns out to be V. Many other functions are ruled out by the same limitation. For example, it is not possible to build a finite-state machine that reads a sequence of binary digits and determines whether the number of 1's is equal to the number of 0's. Similarly, although a finite-state machine can add binary numbers, it cannot multiply them; I leave it to the reader to deduce why.

Beyond finite-state machines and regular languages there extends a hierarchy of more powerful machines and more general languages. It is called the Chomsky hierarchy, after the linguist Noam Chomsky, who investigated the various formal languages as possible models of natural language. The more general languages are created by relaxing constraints on the grammatical rules of regular sets; the machines are built by adding memory elements to the basic finite-state model.

The next machine in the series is called the pushdown automaton. It consists of a finite-state machine with the addition of a memory array that has an infinite capacity but a peculiar organization. The memory takes the form of a stack, like a counterweighted stack of cafeteria trays. An item of information can be stored only by putting it on top of the stack; when the information is retrieved, any overlying items must first be removed. Thus the last item in is the first one out.

The language recognized by a pushdown automaton is called a context-free language. In parsing its statements the acceptability of a symbol can depend both on the symbol immediately to the left and on the one immediately to the right. This bidirectional dependency is permissible because any symbols whose interpretation cannot be decided immediately can be stored on the stack until the ambiguity is resolved. Hence a pushdown automaton can work with subtractive Roman numerals, and it can identify expressions with equal numbers of 1's and 0's (or other symbols, such as left and right parentheses). On the other hand, it cannot detect sentences with equal numbers of three symbols, such as 0's, 1's and 2's. Most programming languages are context-free, and the parser of a compiler is generally a pushdown automaton. Many computers include hardware facilities for organizing a part of the memory capacity as a pushdown stack. One programming language, Forth, makes a stack the primary memory structure. Of course, the stack in any real machine cannot have infinite depth.

The context-free languages merit their name because the parsing of any symbol can be influenced directly only by the symbol's two immediate neighbors, not by the wider context in which it is found. Removing this constraint gives rise to a context-sensitive language and once again increases the difficulty of interpretation. Now widely separated symbols can interact; in the worst case it is not possible to interpret the first symbol in an expression until the last one has been read. In exchange for the added complexity somewhat greater capability is gained. A machine based on a context-sensitive language can determine whether an expression includes equal numbers of three symbols.

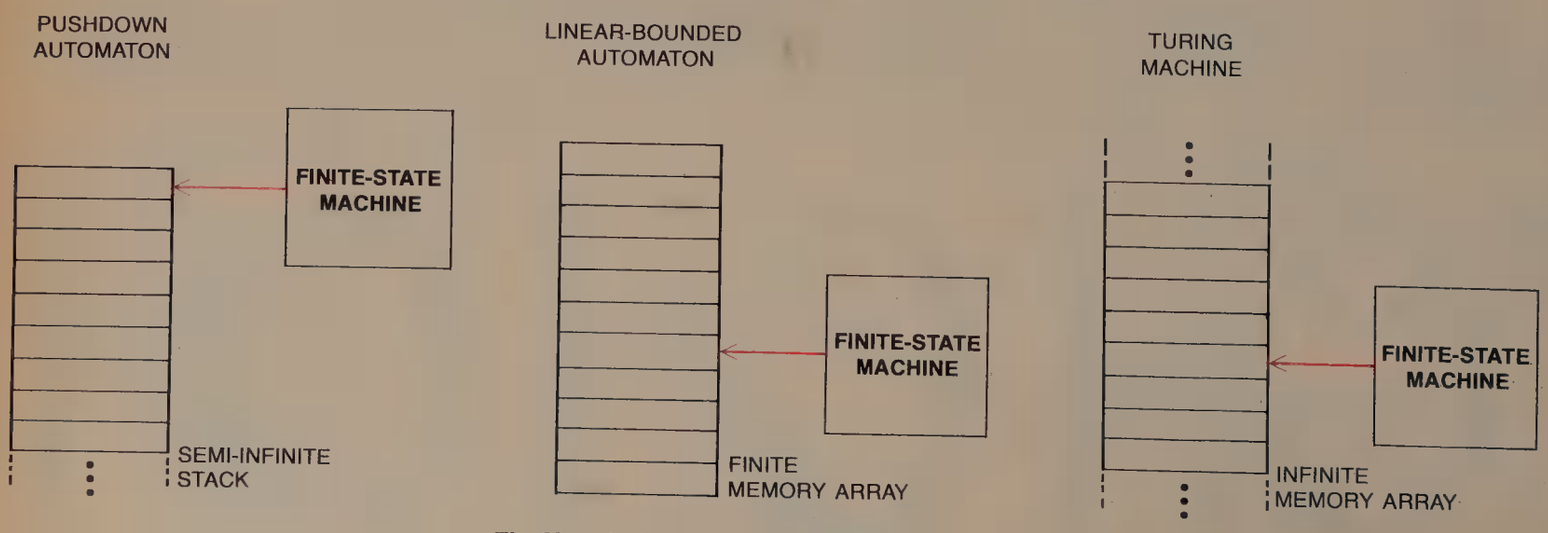
The machine that can recognize a context-sensitive language is a linear-bounded automaton. In addition to the usual finite-state apparatus it has a memory organized in such a way that any storage location can be reached at any time; it is a random-access machine. The memory is only finite in capacity, but it is assumed to be large enough to hold any input the machine receives. The linear-bounded automaton seems a good approximation of the von Neumann model of a digital computer. Oddly, though, the corresponding context-sensitive programming languages seem to be rare; evidently the simpler context-free structure almost always has sufficient expressive power.

All the languages described above have a property in common: they are said to be recursive. What this designation amounts to is that one can imagine a procedure for generating all possible "utterances" in the language in order of increasing length. It follows that there is a guaranteed method of deciding whether any given statement of finite length is a member of the language: simply generate all the statements up to that length and compare them.

There are languages that cannot meet even this minimal standard of tractability. For them there is only one possible recognizing machine: it is the computer of last resort, the Turing machine, a finite-state automaton allowed to roam freely through an unbounded memory. In the description given by Turing the memory is a tape, infinite in both directions and marked off into cells, which the finite-state apparatus can write on, read or erase.

Looking down from the elevated perspective of the Turing machine, the relations of the lesser computing devices become clearer. The linear-bounded automaton is simply a Turing machine with a finite tape. The pushdown automaton has a tape that is infinite in one direction, but the "head" for reading and writing on the tape always remains fixed over the last nonblank cell. The finite-state machine is a Turing machine with no tape at all.

Brand-name-conscious readers, eager to parse nonrecursive languages, may already be out shopping for a Turing machine. They should be warned that the ultimate computer also has its weaknesses. There are languages with grammars so preposterous that even a Turing machine cannot be counted on to recognize their statements in a finite amount of time. So far such languages have found little use in the world of computing machines, but people somehow manage to speak them.



The Chomsky hierarchy of finite and infinite machines

"The center of a literary and political squall."

— The New York Times

"An extraordinarily readable journal of ideas."

— Fortune



Oleg Prokofiev on his father's music • A. L. Rowse on Edmund Wilson • Leonid Pasternak on the Pasternak family • The diaries of Roy Fuller • Malcolm Cormack on the art of George Stubbs • E. M. Cioran on himself • Lewis H. Lapham on knowledge • Rosalyn Tureck on Bach • Mark Girouard on Victorian mansions • Richard Brookhiser on the art of Igor Galanin • Peter Witonski discovering the writings of Bruno Schulz • William F. Buckley, Jr. on our fading schools • A story by Andrei Platonov • Jean MacVean uncovering the poetry of Thomas Blackburn • I. E. Ouvaroff on the fable of Marc Chagall • Bryan Griffin on the Philistines • Otto J. Scott on servility • Martin Seymour-Smith on England • Curtis Cate on the follies of Walter Lippmann • A. L. Rowse on Cosgrave's Dublin • Lev Navrozov exposing the *New York Review of Books* • Richard Brookhiser scolding the *Nation* • Lev Navrozov exposing the *New York Times* • John Chamberlain reminiscing about journalism • A novella by Mircea Eliade • George Gilder on money and freedom • Roy Strong on Nicholas Hilliard and the English miniature • The letters of Arthur Schnitzler • I. E. Ouvaroff on the paintings of William Bailey • A forgotten play by Eugene Schwarz • J. Jean Aberbach on his friendship with Fernando Botero • Uwe Siemon-Netto on the fall of Europe • Sir Ernst Gombrich on Franz Schubert and the Vienna of his time • Lewis H. Lapham on ignorance • Annie Dillard on faith • Ezra Pound on music • Vasily Rozanov on himself • Thomas Molnar on Jean-Paul Sartre • I. E. Ouvaroff on Will Barnet • Gordon Craig's Paris Diary • Boris Goldovsky on opera • William French on Pound's Hemingway connection • Luigi Barzini on the British • Homage to Elie Nadelman • Curtis Cate on Le Corbusier • Raël Jean and Erich Isaac on the utopian think tanks • Also Philip Larkin, W. Nelson-Cave, Charles Causley, Charles Edward Eaton, Joseph Brodsky, Douglas Dunn, Hamish Guthrie, J. C. Hall, John Heath-Stubbs, Leslie Norris, Anne Ridler, Rudolph Schirmer, Michael Schmidt, Tarjei Vesaas, Christopher Fry, Eugene Dubnov, Richard Eberhart.

When *The Yale Literary Magazine* was founded, Beethoven was completing the *Missa Solemnis*, Coleridge's *Biographia Literaria* appeared, and Emerson began his *Journal*. The names of a few authors we have published since then — Rudyard Kipling, Sinclair Lewis, Stephen Vincent Benét, Thornton Wilder, John Dos Passos — show that some of our judgments have been quite timely.

Today, as ever, it is unique talent, not just prominent names, that we seek. When we publish Lewis H. Lapham, or E. H. Gombrich, it is not because one is the editor of *Harper's* and the other a highbrow "name"; what they contribute to our pages is unique, and that is the reason for their inclusion. The same is true of all our authors.

The Yale Literary Magazine still stands alone.

Our magazine is not published fortnightly on newsprint, to extol this week's writers, expound on last week's thinkers, and crumble to dust a week thereafter. The paper we print on permits the most accurate color reproduction of any magazine in the world and is guaranteed to endure for centuries. And we seek to publish and reproduce what will last at least as long.

But *The Yale Literary Magazine* is now under attack.

So boldly, so uncomprisingly have our authors spoken on issues of culture and politics that the academic tastemakers of the university in whose shadow the magazine was born have sought to suppress it as a source of intellectual dissent. Their actions, as George Will remarked in a recent **60 Minutes** broadcast dealing with the controversy, have cast doubt on "the integrity of a major American university," and compelled the magazine to defend its freedom in court.

Our freedom endures, and *The Yale Literary Magazine* remains as independent, as passionate, as controversial as when it first addressed the nation. In the words of reviewers, it is "elegantly produced" (**The Washington Post**) "in the fine book tradition" (**Folio**), "sophisticated" (**American Spectator**), "impressive" (**Los Angeles Times**), "highbrow" (**Time**). "This spendid journal" (**Anthony Harrigan**), "strikingly handsome" (**Chronicle of Higher Education**), "downright lovely to look at" (**James J. Kilpatrick**), is also "good, mean fun" (**Los Angeles Times**). In short, *The Yale Literary Magazine* is "an organ of the *intelligent* intelligentsia" (**Eugene V. Rostow**).

Join us, and see for yourself why "*The Yale Literary Magazine* is one of the few magazines now being published in the United States that attempts a serious encounter with literature, with thought, with ideas."*

The Yale Literary Magazine

Since 1821

The Yale Literary Magazine is a publication of American Literary Society, Inc., a non-profit organization.

*Lewis H. Lapham

Order Form: Introductory Rates

☐ I wish to subscribe for ☐1 ☐2 ☐3 year(s) at $24 a year ☐ Check box if this is a gift.

Name _____
(Please Print)

Address _____ Apt. _____

City _____ State _____ Zip _____ Country _____

☐ My check for $ _____ is enclosed ☐ Bill me

Checks are payable to:
The Yale Literary Magazine
Box 243-A Yale Station
New Haven, Ct. 06520

Please charge my ☐ VISA ☐ MC ☐ AMEX
Card # _____
Expiration Date:
For faster service call (203) 624-8400.

DEF 45

BOOKS

A Christmas survey of books on science and technology for the younger reader

by Philip and Phylis Morrison

This year we found before us a large number of computer books. Most of them turned out to be so specific as to be mere instruction manuals, useful and even necessary but lacking general interest. We hope that deeper treatments will come. Meanwhile we present a collection of other topics, this time richer for older young readers than for the youngest ones.

Laboratory and Field

FINDING YOUR WAY ON LAND AND SEA: READING NATURE'S MAPS, by Harold Gatty. The Stephen Greene Press, Brattleboro, Vt., and Lexington, Mass. ($8.95). A man or woman "with a good sense of direction is, to me, quite simply an able pathfinder... who can find his way by the use of the five senses... he was born with... developed by the blessing of experience and the use of intelligence." Harold Gatty wrote this little classic of pathfinding without map or compass in his Fiji home not long before his death; it was first published in 1958. A ship's officer and a pioneer in aerial navigation, he became a famous air navigator on the occasion of his 1931 round-the-world flight with Wiley Post. He maintained a lifelong interest in finding one's way without benefit of costly gyroscopes. The book conveys the quality of an expert not by book alone but at first hand. It includes a learned look at surprising ethnographic material.

The smell of burning aspen in the Rockies usually means a homestead; the smell of pine smoke might mark a forest fire. Burning peat signals the Falklands many miles at sea; the Cape Verde Islands waft the scent of orange groves afar; the odor of drying coconut often carries out of sight of the atoll whence it comes; seabird colonies broadcast the ammoniac guano smell, welcome enough to a lost sailor. Brigadier R. A. Bagnold, a gifted student of the desert, once found a water hole in broken country "entirely by the smell of one camel, which was picked up eight miles away." In the lifeless desert the background of organic smells is low: the signal-to-noise ratio is good.

Gatty's point is less the utility of our underused sense of smell than it is the value of meticulous attention to local circumstances. He develops the theme brilliantly. Moss on the north side of trees is only the most banal of a dozen different clues from the habit of trees and other plants and from the farmhouse too. These clues are regional rather than general: local knowledge counts. The moss clue is a "dangerous generalization" in the old books. The preferred direction depends on wind and rain; moisture is what the mosses seek. In the Jura Mountains mosses and lichens mark the southwest. Termite nests, ridges on polar snowfields and the form of desert dunes offer more exotic orienting data and exemplify the method.

Only the Micronesians and Polynesians have understood how to read the ocean swells for signs of distant land. One of their wave-pattern diagrams woven in coconut palm is shown. There is a fresh small chapter on ethnocartography; quite a few records exist of the use of maps, both mental and graphic, in a good number of preliterate cultures, often in training contexts. There is even a chapter on finding your way in towns: television aerials, shutters, rust, worn paint, traffic gradients—also excellent training. Distance? Fifty yards is the distance to a man when you first see the white of his eyes, as it was at Bunker Hill; at 800 yards a man looks like a post. An average church steeple can be made out at nine miles.

There are now other books of this ingenious art, but these inexpensive words, with sun tables, star maps and bird pictures, all by a master, are welcome. Swift, lively reading, Gatty should be prized by many a young adventurer out in the woods or in an armchair, from the upper grades onward.

VOLCANO WEATHER: THE STORY OF 1816, THE YEAR WITHOUT A SUMMER, by Henry Stommel and Elizabeth Stommel. Seven Seas Press, Inc., Newport, R.I. ($15). You won't find it in *Old Farmer's Almanac.* Contrary to folk legend, this venerable source of forecasts said not a word about the unseasonable frosts that struck the New England highlands in all 12 months of the year 1816. The facsimile pages are here to prove it. That year deserved its folk name of "eighteen hundred and froze to death."

In this delightful example of science and scholarship done persuasively as one the Stommels (he is a veteran oceanographer from Woods Hole) marshal the evidence to make clear what happened in the strange summer of 1816. From primary evidence, notably the meticulous temperature records kept by a network of college presidents, including those of Harvard and Yale, it is plain that the summer went awry. New Haven had the mean June temperature we normally expect at an inland site 200 miles north of the city of Quebec. The frost-free period at Yale lasted for only 71 days that summer, although the mean value in a cool decade was 126 days. Those hardscrabble northern Yankee farmers who depended on the maize they grew as a staple had very poor crops, and in interior Quebec even the hardier wheat failed. The farmers of Maine and Vermont set out for Ohio and Indiana in droves; although the emigration was not, of course, solely in that year or solely for the reason of climate, graph after graph supports the idea that 1816 was a year of flight.

The probable cause is known: dust in the high air from a remote eruption. The volcano Tambora exploded in April, 1815, depositing ash a foot thick over the island of Bali a couple of hundred miles to the west. It is thought to have been the greatest volcanic explosion in historic times, the volume of ejecta 100 times those of Mount St. Helens and 10 times those of Krakatau. In a fine closing chapter the reader is offered the summary data, in aligned curves over the decades for sunspots, temperature and sources of volcanic dust, and is challenged to form an opinion. Formal statistics help little; the authors, and at least these readers, see the data as suggesting that volcanoes cause cold spells. Whether dust cooling will favor us by compensating for fuel-fed carbon dioxide warming, or whether the soots of nuclear-ignited cities and forests will spell worldwide catastrophe, are questions the future will decide. Any able young readers—or their parents—can learn to appreciate and to share the concerns from this brief, engrossing and scrupulously honest work.

KITES: THE SCIENCE AND THE WONDER, by Toshio Ito and Hirotsugu Komura. Japan Publications, Inc. Distributed by Kodansha International/USA, Ltd., through Harper & Row, Publishers, Inc. ($11.95). Old Japan: the strong, subtle handmade paper of mulberry and the light, tough, elastic lengths of bamboo are worked into kites of high performance and intriguing shape in the hands of masters working in a tradition of centuries. New Japan: over four years

THE GIFT OF LOVE



AMARETTO DI SARONNO ORIGINALE·1525

TO SEND ANYWHERE IN THE U.S. CALL (800) 528-6148. VOID WHERE PROHIBITED.

LIQUEUR 56 PROOF. SOLE U.S. DISTRIBUTOR, FOREIGN VINTAGES, INC., NEW YORK, N.Y. © 1983. A UNIQUE PRODUCT FROM ILLVA SARONNO ITALY



The new Mercedes-Benz 380 SE. With its 3.8-liter light-alloy gasoline V-8, sophisticated fully independent suspensi

The performance renaissance accelerates: introducing the new Mercedes-Benz 380 SE Sedan.

THE NEW 380 SE joins the Mercedes-Benz line at a most propitious moment–just as America rekindles its appetite for mightily responsive automobiles.

An automobile more likely to satisfy that appetite is difficult to conceive.

The 380 SE is, in a word, potent. It combines a 3.8-liter gasoline V-8 engine with a 3,685-lb. curb weight, creating a ratio of power to weight conducive to formidable over-the-road performance.

As invigorating as the 380 SE's turbine-smooth acceleration is its willing throttle response at almost any speed. Power reserves seldom feel less than massive. Yet in stop-and-go urban traffic, the Mercedes-Benz 380 SE feels as docile as the proverbial lamb.

This robust V-8 engine is testimony to how far Mercedes-Benz has advanced performance-engine technology since the age of the cast-iron behemoths.

A LIGHTWEIGHT V-8

For example, the block itself is one of numerous engine components fabricated in *aluminum alloy*–making this V-8 as notable for the weight it saves as for the power it generates.

It is meanwhile as resistant to heat, vibration and wear as a conventional iron-block V-8. Note that the pistons ride on cylinder walls having a silicon surface–a substance so hard that it must be machined by diamond-tipped tools.

Mercedes-Benz engineers introduced fuel injection to passenger cars. In the 380 SE, fuel is metered to the cylinders by a sophisticated third-generation C.I.S. continuous injection system.

Power flows to the rear wheels through a crisp-shifting four-speed torque converter automatic transmission.

The 380 SE's performance is maximized by its extremely low wind resistance. (Why battle the wind, when you can outwit it?) Its 0.37 aerodynamic Cd comes within a whisker of the most ballyhooed aerodynamic sensations of the day.

RESPONSIBLE ROADHOLDING

The 380 SE's forged light-alloy wheels are the most visible elements of a suspension system set up for determined driving.

Mercedes-Benz believes that a performance automobile that was all speed and no handling would be an act of engineering irresponsibility. Thus, you may be as taken with the 380 SE's roadholding as with its power.

You may be equally taken by its steering precision. A power-assisted recirculating-ball steering system lets you guide this substantial machine with pinpoint accuracy. No luxury-car sponginess.

*See your authorized dealer for complete warranty information. ©1983 Mercedes-Benz N.A., Inc., Montvale, N.J.



ed four-wheel disc brakes, this is one performance sedan meant to go, to handle, and to stop.

Automobiles built to move this well should also be built to stop with equal responsiveness. The 380 SE is engineered to stop with the power of four disc brakes. Total swept brake area: 456 square inches.

CRAFTSMANSHIP LIVES

The cabin accommodates five persons within 93.6 cubic feet of volume. Their comfort is further enhanced by an array of electric, electronic and other amenities. Especially ingenious are the 10-way electric adjustment controls for each front seat–activated by a tiny facsimile of a seat.

The 380 SE confirms, among other truths, that patiently crafted automobiles are still being produced in the 1980s. It is an example of consummate workmanship–from the precise fit of every panel to its hand-finished Zebrano wood interior trim.

Mercedes-Benz wants your 380 SE to be as pleasurable to own as it is to drive. One proof of this is a new *48-month-or-50,000-mile* limited warranty.*



RESALE PERFORMANCE

Mercedes-Benz automobiles have been shown, year after year, to retain a high percentage of their original value. This performance can be as reassuring, in its way, as the 380 SE's over-the-road performance.

The ultimate 380 SE accolade may have been voiced by the Editor of *Car and Driver,* in a recent survey of America's most desirable automobiles: "There is nothing like a big Mercedes V-8. Nothing. It is a car that is simultaneously fun to drive, strong as a D-8 Cat, and luxurious enough for all but the irresponsibly sybaritic."

Engineered like no other car in the world



THE FISH WON'T BE BITING IN TWITCHELL CREEK TODAY,

OR TOMORROW, EITHER.

There was a time when you could pull a fair-sized fish out of Twitchell Creek on Woods Lake in the Adirondacks. But no more.

The fish are gone . . . along with many other creatures who once lived in the water and along the banks. They've been run off or killed off by the rain. Deadly, poisonous acid rain which has choked the life out of the stream, and the land as well.

Acid rain is deadly because it contains nitric acid and sulfuric acid. These killers are born when rain water mixes with nitrogen oxide and sulfur dioxide. Two chemicals which are being spewed by the ton-load into our air every day by coal-burning power plants and industrial boilers across our land.

Some lakes contain neutralizing "buffers" which lessen acid damage. But what of the others . . . in the Adirondack Mountains, in western Virginia, throughout New England . . .

Who Will Stop The Rain?

The Izaak Walton League, that's who. Formed in 1922 by a handful of conservation-minded sportsmen, the League is now 50,000 strong . . . and ready to fight those who want to relax the 1970 Clean Air Act regulations which set allowable limits for sulfur dioxide emissions. We want stricter laws. And we'll get them. Because there's something at stake here far greater than fishing. And that's life itself.

For more information on our activities write:

THE IZAAK WALTON LEAGUE OF AMERICA

Izaak Walton League
1800 North Kent Street
Arlington, Virginia 22209

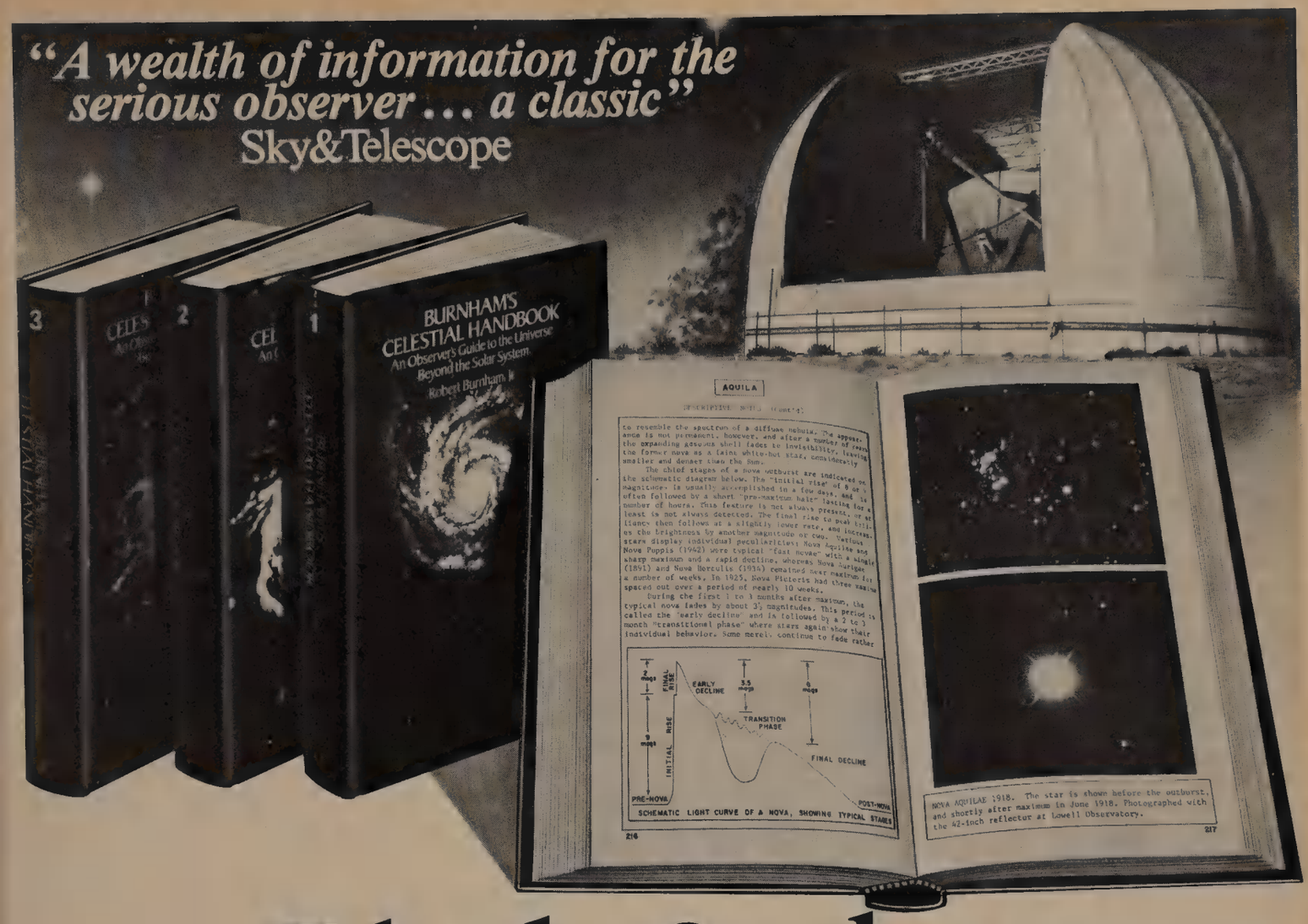
two expert engineers, making and flying together more than 1,000 trial kites, develop a full-blown aerodynamic theory of kites and elaborate a system of trial design, accessible to earnest novices, achievable in materials both old and new, high in performance and shaped after birds, insects and fishes. This is their own account, published in 1979 in Japan and newly translated into serviceable English.

The opening chapter, with many force diagrams and performance curves, puts before a reader who has some grasp of high school mechanics the most complete theory of kites in stable flight yet published. That is followed by a well-argued account, although one much less complete and quantitative, of kites seeking stable flight: the spins and yaws and flapping that plague every builder. The latter half of the brief but full book is less analytic: it looks to construction. Materials are discussed and their properties tabulated; modern wonders such as Tyvek and graphite composites are included, although not nylon sailcloth, a staple of U.S. kite builders. The authors' system of design and construction is outlined, and detailed kite forms are drawn and discussed. A terse final chapter treats wind and the kite builder's adjustments in response to it.

In an empirical art this old and refined the rise of a full theory, to which the largest contribution has been made by these two authors, is not perhaps an obvious event. Their kites perform very well in a free choice of forms, but so do many others. No serious kite builder can afford to overlook this book; it opens an entire road of investigation for clubs and informal teams of kitepersons, whose patience and intuitive skills can now be turned to rewarding engineering at the level of high school physics and well beyond.

If the reader is not misled, the implicit scaling of these formulas suggests that kites built on plausible rules of proportion will fly in wind speeds that decrease with the diameter of the kite only as the square root, and that the altitude the kite can reach under the load of drag and string weight increases in the same slow way. The sky is forgiving to both small kites and large, structure permitting.

THE SCIENCE IN SCIENCE FICTION. General editor, Peter Nicholls; contributors, David Langford and Brian Stableford. Alfred A. Knopf, Inc. ($25). Nearly 100 thoughtful pieces of only two pages each are assembled into related topics by chapters and are supported by a wide choice of illustrations, most of them, although not all, reproduced or redrawn from published sources. The themes are part and parcel of the received doctrines of science fiction: starships, aliens, hyperspace and hyperspeed, time travel, nuclear holocaust,



Take the 3-volume Burnham's Celestial Handbook for only $3.95

when you join the Astronomy Book Club

a $60.00 value

You simply agree to buy three more books—at handsome discounts—within the next 12 months.

The most complete sky manual ever. Three easy-to-use hardcover volumes (a $60.00 value) guide amateur and professional astronomers to a greater knowledge and enjoyment of the stars. Detailed coverage of 7,000 deep-sky objects that are accessible to observers with telescopes in the 2″ to 12″ range. Provides names, celestial coordinates, classification and complete physical descriptions of almost every deep-sky object.

Entries are grouped according to constellation with extensive descriptive notes covering historical and mythological associations, unusual movements and appearances, plus explanations of white dwarfs, novae and supernovae, dark and gaseous nebulae and hundreds of other topics. Numerous charts, diagrams and graphs included to help in identification. More than 2,000 pages containing over 600 photographs.

"Totally professional and competent in virtually every respect, *The Celestial Handbook* is for anyone who follows the stars."—*Star and Sky*

Astronomy Book Club is a very special club—a book club operated exclusively for astronomy enthusiasts. It's devoted to making the finest astronomy books available conveniently and inexpensively. Recent offerings include titles by such eminent authorities as Carl Sagan, George Abell, Robert Jastrow, and John Gribbin.

As an **Astronomy Book Club** member, you'll receive books of your choice at savings up to *30% off publishers' prices*, and occasionally even more. In addition, you'll immediately become eligible for our Bonus Book Plan, with savings up to 70% off the publishers' prices.

Discover the phenomena of space—join the **Astronomy Book Club** today.

If reply card has been removed, please write to: **Astronomy Book Club**, Dept. L-AX8, Riverside, N.J. 08075 to obtain membership information and an application.

4 Good Reasons to Join

1. The Finest Books. Of the hundreds and hundreds of books submitted to us each year, only the very finest are selected and offered. Moreover, our books are always of equal quality to publishers' editions, *never* economy editions.

2. Big Savings. In addition to getting BURNHAM'S CELESTIAL HANDBOOK for only $3.95 when you join, you keep saving substantially—up to 30% and occasionally even more. (For example, your total savings as a trial member—including this introductory offer—can easily be over 50%. That's like getting every other book free!)

3. Bonus Books. Also, you will immediately become eligible to participate in our Bonus Book Plan, with savings up to 70% off the publishers' prices.

4. Convenient Service. At 3-4 week intervals (16 times per year) you will receive the Book Club News, describing the Main Selection and Alternate Selections, together with a dated reply card. If you want the Main Selection, do nothing and it will be sent to you automatically. If you prefer another selection, or no book at all, simply indicate your choice on the card, and return it by the date specified. You will have at least 10 days to decide. If, because of late mail delivery of the News, you should receive a book you do not want, we guarantee return postage.

robots, clones, telekinesis, flying saucers and force shields. The three British authors bring impressive credentials. The general editor is a master of the canon of science fiction; the two contributors were trained respectively as a physicist and a biologist and are both science-fiction writers as well. The illustrations include reality, such as the jet of the radio galaxy M87 and a wide defoliated strip along the DMZ in Vietnam; travesties, such as Erich von Däniken's slightly retouched version of a Mayan relief seen as a drawing of an astronaut; art, such as the bronze she-wolf suckling Romulus and his brother, and plenty of fictional representations, from the great mother ship of *Close Encounters of the Third Kind* to the less ethereal covers of old issues of *Amazing Stories.*

In just about every piece there is a literary mention of the science fiction of the past and present that includes that particular theme, followed by a popular but knowing account of the actual situation as far as scientific knowledge or likely consistent conjecture goes. The entire volume stands as clear evidence of how pervasive the imaginative ideas of science fiction in fact are, of how many of the genuine dilemmas of a world immersed in the implications of a newly powerful science are reflected in the popular genre in all media, sometimes intelligently, sometimes merely for calculated effect.

The next most striking fact is that all these discussions are sensible, even helpful, on their straightforward expository level. Perhaps not everyone would agree with all the conclusions, since they range over human expectations broadly. Not one of the brief essays, however, seems in scientific error or illogical; quite the contrary. A deal of best-selling junk is disposed of by the evidence, and plausible conjecture is compared with what we know and what we only surmise. For example, light speed and time travel, even black holes and tachyons, gain entirely reasonable exposition here.

The book is a bridge between fiction and reality that does its fiction-oriented architects proud. Young readers from the middle grades on would find it absorbing; school libraries need it at any level, a sure winner for topical reference and bibliography, fact as well as fiction.

KITCHEN CHEMISTRY, by Robert Gardner. Illustrated with photographs and with line drawings by Jeff Brown. Julian Messner/Simon & Schuster ($9.29). Stove, sink, counter and refrigerator, the household kitchen is a well-equipped laboratory, better than many a college student can boast for a good deal of chemistry and physics. Using it requires tact, since the house depends on it. Given a degree of thoughtfulness, any young person can find satisfying results in the kitchen.

This small book outlines an entire set of experiments for each of the major pieces of kitchen equipment. At the sink it is capillarity that draws first attention. Bubbles, gas collection and jets are not forgotten, even the electrostatic attraction of a water jet (perhaps not as easy as it looks). Melting ice and boiling water offer a chance (it takes patience) to follow in the famous footsteps of Joseph Black, who first understood and measured those latent heats by simple timing. Electrolysis is proposed safely, relying on battery power. There is a good deal more, all of it solidly workable. This is no mere stew of plausible old proposals but a sound guide to what actually works. (Candles burned in sealed jars go out before they use up all the oxygen in the air, and the way to prove it is here.) The enthusiastic book is overrich; it would take months for a devoted kid to carry out all that is in it.

RAIN SHADOW, by James R. Newton. Illustrated in black and white by Susan Bonners. Thomas Y. Crowell ($10.95). The book begins with a view at the shore, the air heavy with the ocean moisture. Clouds form and float above the evergreen forest; we are in the Cascades. In the high cold clouds the droplets of water bump, grow larger and fall as rain on the mountain slopes nearly every day of the year. Here are ferns and mosses, tall trees draped with green and bushes growing from every decaying log; even the air seems green. We climb to flowery snow-fed alpine meadows, and then to the snowfields themselves, too cold for much life. Now the air has lost its water; the valley beyond lies in a rain shadow. Cactus and sage abound; the farmers can grow crops only because irrigation canals capture some of the river water that flows into the valley from the heights. The case is not unique; a few others are mentioned, where ocean-born rainfall is shadowed by mountains. The clear text, a few sentences to a page, not without useful detail and even a few numbers, is strengthened by the textured and convincing charcoal drawings. The book is a cool pleasure for readers in early grades who like to travel in reality or in imagination.

Plants and Animals

DINOSAURS: AN ILLUSTRATED HISTORY, by Edwin H. Colbert. Hammond Incorporated ($30). THE SUCCESSFUL DRAGONS: A NATURAL HISTORY OF EXTINCT REPTILES, by Christopher McGowan. With original illustrations by Marg Sansom-Markezinis. Samuel Stevens and Company, Toronto and Sarasota ($29.95). Finicky taxonomists fluent in a Hellenic jargon of their own, the vertebrate paleontologists in the dinosaur trade remain admirably broad-minded, lively and articulate scientists. They can write attractively for the general reader, and they do. Here are two top specimens of the art.

Dr. Colbert, for 40 years responsible for the exhibit halls that display these large fossils at the American Museum of Natural History in New York, is dean of American dinosaur experts. Dr. McGowan is for his part curator of the splendid renewed displays at the Royal Ontario Museum in Toronto; the books are manifestly both authoritative. The Colbert book offers more of what people call production value: its pages are full of colorful reconstructions, some classical, some new. Its opening spread presents a striking color close-up of the Victorian reconstruction of the first dinosaur recognized, the iguanodon, its image made of tile and painted concrete for the Crystal Palace near London in 1854.

McGowan's book, handsomely illustrated without quite such panache, begins instead with a brief survey of all the orders of reptiles, and it treats of Darwin as historical prelude. McGowan is a specialist in the marine and the airborne reptiles of the past, not strictly speaking dinosaurs at all, so that his book takes a wider look at variety, with its center of gravity in the physical and biological problems of flight, of swimming fast at sea and of the scaling constraints on being a giant. Colbert instead centers his text and pictures on a winning account of the variety of the creatures—all those saurians, large and small, track, egg, skin and bone. McGowan tells, say, of the Reynolds number, a remark of a kind not to be found in Colbert; Colbert details many forms that do not occur in McGowan. The vivid books share an air of sustained high quality; one might see the Colbert, plainly intended for a wide family audience, as a comprehensive introduction to the entire topic, whereas the McGowan, more text than image, is a second book on ancient reptiles for those who seek a little more connection with biology as a whole.

Like evolution itself, the books converge a good deal. Both discuss very well the issues of the warm-blooded and vigorous behavior of some, not all, of the fossil reptiles. Both agree that the easy answer of asteroid collision, or indeed any other known cause, cannot really explain the rapid extinction of the beasts. (Indeed, these writers both think the true cause may be unknowable, but we shall see.)

One image might unite the books. The Red Deer River flows southeastward through the little Alberta town of Drumheller, now the "focal point of the dinosaur cult." The sandstone badlands of this part of Alberta have already yielded about 500 specimens, more perhaps than any other place in the world. New finds weather out of the soft rock formations year after year. The idyll

(mosquitoes apart) of this wonderful vocation occurred in the early years of the century, when the collectors drifted down that cool river on long lazy summer days, scanning the high banks from big tented rafts, to find fossil after remarkable fossil. Colbert shows one such floating field camp in a period photograph, and McGowan puts an end to the period, if not to the narrative, in a photograph that shows a thin band of coal marking a great change: below it a world ruled by reptiles, above it one under mammalian sway. At Hell Creek, Mont., an observer can stand with his feet in one era and his head in the other, a world revolution in at most a couple of million years.

WALKER'S MAMMALS OF THE WORLD, by Ronald M. Nowak and John L. Paradiso. Fourth edition, Volume 1 and Volume 2. The Johns Hopkins University Press ($65). The first edition of this grand work of reference appeared in 1964, the outcome of 30 years' work by Ernest P. Walker, then assistant director of the National Zoological Park. His expert continuators have brought the work up to date in a series of new editions; here it is, better than ever, the only reference that offers information on just about every known genus of the mammals, our own class. More than 4,000 of these marvelously diverse species are included, about 1,000 genera, in 1,400 pages.

Plainly there is not much space available for any single species, even for human beings, who are given two and a half pages but no pictures. That so many forms are pictured is the strength of this comprehensive reference for young readers and their libraries. The text is moderately technical, with terms of art for structure and color. The brief articles include as a matter of course, however, the distribution, relationships and interesting behavior for all the genera and most individual species. There is a black-tailed prairie dog (no dog, but a stout, short-legged squirrel) alert in its conical crater, and the reminder that many years ago in western Texas there was a large colony, a town, of the animals said to have sheltered 400 thousand individuals. Here is a shelly three-banded armadillo from Brazil, curled up in Walker's hand and resembling a coconut. There is the whitefin dolphin, which lives along the Yangtze River for 1,000 miles up from the mouth; it has long enjoyed protection by custom, and it is now under legal protection as well, although silting, motorboat collision and water use are reducing its numbers.

It is up to the reader, however, to seek out more interesting stories among the thousands here. A table of distribution among world regions covers every recent genus of mammal, and an indispensable index is found in each volume, with both common and Latin names. More than 3,000 papers and books are cited in detail. Less showy than its authoritative treasures merit, Walker's is a reference bargain for any library where young readers and old enjoy animals.

OAK & COMPANY, by Richard Mabey. Illustrated by Clare Roberts. Greenwillow Books, William Morrow & Co., Inc. ($9.50). "One winter night in its 283rd year, the Oak was hit by a blizzard." The old tree had only a few leaves as more and more branches were shaded out; its weakened roots could not hold the heavy load of snow and the great tree toppled. The foresters came soon and cut it up; the timber was of little use with its crooks and hollows. The tree rotted slowly away, still host to toadstools and insects, as in life it had held a great company, woodpeckers and butterflies, mosses and beetles, now moved to the straight trunks of the Oak's offspring nearby, timber oaks perhaps in the far future.

In this beautifully illustrated book young readers will share the long, slow story of an oak, a living organism that begins as an acorn, matures to bear its own young only in its 49th year, then for a long time endures chance and the elements. Human beings visit and use this English forest thoughtfully over centuries, and an entire community of life centers on and under the great tree, including a million earthworms slowly eating through the litter of the forest floor. The pace is slow and majestic, although we move also to the swifter rhythms of woodpeckers and pigs, bluebells and cepes, coppiced stumps and spangled galls. The quiet account (with a few phrases unfamiliar in American speech) would appeal to thoughtful middle-grade readers.

THE CRANBERRY BOOK, by Elizabeth Gemming. Coward-McCann, Inc. ($9.95). If the Royal Navy banished scurvy by issuing West Indies limes to its men, the limeys had a counterpart in American ships. The whalers and the smart China clippers carried tart red berries, rich in vitamin C, able to stay fresh for two years packed in barrels filled with spring water. This is a well-written and interestingly illustrated little monograph on the cranberry, suited for the right readers in the middle grades, even though it tells more than many might care to know.

Today Massachusetts and Wisconsin grow more than four-fifths of the North American crop. The term cranberry bog is a word of art: cranberries grow well only in drained places, not in true wet bogs. The roots require the acid peat subsoil of a former bog; the runners thrive in a layer of sand; the crop is typically watered by sprinklers. Nowadays the berries are harvested by flooding the ripened vines, churning the shallow water mechanically into turbulence so that the berries come off their stems to float "like a gorgeous ruby carpet."

The Wampanoag used the wild berries well, but there is no record of cranberry sauce at that first Thanksgiving of 1621. Cranberries were a wild resource of no value until Captain Henry Hall of East Dennis first transplanted them in about 1816. The national crop is today some two million barrels per year (a small barrel, peculiar to cranberries, of less than 25 gallons), nearly all consumed as frozen, tinned or bottled products. Visitors to the Cape Cod National Seashore can seek out both wild berries and a small cultivated bog.

ED EMBERLEY'S SCIENCE FLIP BOOKS: THE CHAMELEON/THE CHICKEN; THE FROG/THE HARE; THE BUTTERFLY/THE DANDELION. Little, Brown and Company ($5.95). Six "nature adventures," little dramas of natural history, are presented here in three small, colorful flip books. All the sequenced paintings are carefully and brightly done; for each story there are a few lines of initial text to set the scene, and the rest of some 60 pages are shaped and bound to allow the quick flip that generates a moving picture. The chick pecks out of her egg and grows past a pullet to a hen. The chameleon walks a few slow steps as its color changes. Frog and monarch butterfly metamorphose grandly. The hare's coat turns from winter white to summer brown as it escapes the fox. The dandelion blossoms and sets seed into the winds. Somehow gentle in mood, these hand-sized books are amusing and instructive, whether flipped fast or savored slowly page by page.

FERAL: TAME ANIMALS GONE WILD, by Laurence Pringle. Macmillan Publishing Co., Inc. ($9.95). The rock dove was a wild creature of the cliffs. Buildings of wood and stone offered the birds a new set of cliffs, conveniently near grain and pools of water. By Greek times they were half-tamed. The dovecote housed them at human expense, although they were free to come and go. In compensation the people collected occasional squabs for food. Domestication of the species was clear. Nowadays urban dovecotes are few, and the squabs of the cities are safe. The large pigeon populations are now feral; the mute swan and the monk parakeet, hardy in cold New York City, are other bird species that remain near human settlement, although they have escaped from our direct control.

Bands of feral pigs, omnivorous, fierce and clever, roam the forests of Hawaii, California and the Southeast; there may be two million of them. In the 1870's there were as many as 3.5 million feral horses in the unfenced West. Now

SAVE AT ELEK-TEK

TEXAS INSTRUMENTS

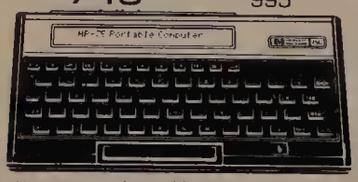
BA-2	33	TI-66	52
BA-35	22	PC-200	52
TI-35-SP	17	TI-5040-2	49
TI-MBA	30	TI-5130	60
BA-55	43	TI-5142-3	75
TI-55-2	33	TI-5218	130
TI-57	29	TI-LCD-Prog	50



HEWLETT-PACKARD

HP-75C 16K PORTABLE COMPUTER

$719 Mfr. Sugg. 995



The battery powered portable computer which includes a built-in HP-IL interface, card reader, and time and appointment modes. Typewriter keyboard.

82164A HPIL/RS-232C Interface 235
82168A HPIL Acoustic Modem 395
82169A HPIL/HP1B Interface 315
82700A 8K Mem. Module 150

NEW! **HP-41CX $245**

The great features of the 41CV

PLUS Built-in:

- Time Module Functions
- Extended Functions Module
- Enhanced Text [ASCII] FILE EDITOR

HP-41C $145
HP-41CV 200



Optical Wand 95
Card Reader 145
Printer 283
Quad R.A.M. [for HP41C] . . **60**
Ext. Memory Module 60
Ext. Function Module 60
Time Module 60

	Mfr Sugg	Elek-Tek
HP-IL Loop Module	125	**95**
Digital Cassette Drive	450	**335**
Printer/Plotter[HP-IL]	450	**335**

SLIMLINE Shirtpocket Styled
Power Packed Programmable
LCD PROBLEM SOLVERS

HP-10C Scientific *(New)* **$54**
HP-11C Scientific 70
HP-15C Scientific *(New)* **90**
HP-12C Financial 90
HP-16C Programmer *(New)* . . . **90**
HP-97 Desktop Scientific . . 560

50%-60% off orig. mfr. suggested prices

HP-32E Scientific 32.50
HP-34C Scientific 50.00
HP-37E Financial 32.50

hp HEWLETT PACKARD | PERSONAL COMPUTATION DEALER

CALL TOLL FREE 800-621-1269
EXCEPT Illinois, Alaska, Hawaii

Corp. Accts. Invited. Min. Ord. $15.00 Mastercard or Visa by mail or phone. Mail Cashier's Check, Mon. Ord. Pers. Check [2wks to clr.] Add $4.00 1st item. **[AK, HI, P.R., Canada add $10.00 first item]** $1.00 ea. add'l shpg. & handl. Shipments to IL address add 6% tax. Prices subj. to change. **WRITE for free catalog. 30 day return policy for defective replacements only. ALL ELEK-TEK MERCHANDISE IS BRAND NEW, FIRST QUALITY AND COMPLETE.**

ELEK-TEK, inc.
6557 N. Lincoln Ave., Chicago, IL 60645
(312) 631-7800 (312) 677-7660

their number is down by a factor of some 200, but what to do about the remnant remains a dilemma. The burros of the deserts and the ownerless cats and dogs that wander city streets are further witnesses to defects in the complex unwritten contracts between people and the few animal species that have come to mean much to human life over the past few thousand years.

Laurence Pringle's simple, informed text will engage more than a few readers in the fifth and sixth grades, whose empathy with these introduced living beings, neither chattel nor yet alien, is not a bad basis for a first encounter with knotty ethical and economic problems. The book is dedicated very affectionately to a Burmese cat we see in one photograph, "so wild, so tame, so fine."

Technology Past and Present

FROM HAND AX TO LASER: MAN'S GROWING MASTERY OF ENERGY, by John Purcell. Illustrated by Judy Skorpil. The Vanguard Press ($16.55). In a deft and personal prose, full of unexpected insights arising from widely shared experiences, the author has made a believable and never credulous narrative of the entire history of human technology. It begins with the beginning, the Rift protohumans of two million years ago, then dwells on the fire users of the cave at Chou-kou-tien. The theme of energy provides a minor framework, if an interesting one, for the account. The simple machines of grade school physics are fully as relevant, and they are very usefully discussed. We find stone knife blades and infer the digging stick, a clear lever. Fire making is pleasantly handled: the conservative archaeologist is shown in doubt about the fragment of charred beechwood rod from a Mousterian site in Yugoslavia. Maybe it was a fire-hardened spear. His radical colleague has no such doubts: it is a friction fire stick. Why not fire making in that cold epoch 100,000 years back? "For myself... I am content merely to hazard that some Neanderthal men knew how to *make* a fire when necessary," Purcell writes.

So the text moves, always lightheartedly but sensible, parting certainty from plausibility, all the way from early hominids as mere scavengers and not mighty hunters to spoked wheels, possibly invented more than once. Up-to-date always, the material used includes the evidence on flint mining in the Neolithic, with the simple reckoning that establishes the large scale of the mining efforts. In the same way we read of the origins of writing, lately shown by Denise Schmandt-Besserat to be linked with the clay tokens commonly found in the preliterate trading villages of southwestern Asia, three millenniums older than any Uruk seals. Pottery, cast metal, wheels for many uses and much more are surveyed in the same tone: the best evidence, its uncertainties and a sensible trial interpretation, without pretensions. The last tenth of the book treats of the heat engine, of James Watt and his context, and on to radio via Reginald Aubrey Fessenden's pioneer voice broadcasts from Cape Cod in 1906. Lasers end the book, a little trendily perhaps, with a few pages on zapping.

The blend of good sense and wide learning, served up with good humor, makes a remarkable book for young people and their elders as well. It is a contemporary introduction to human history and prehistory, not meant to be comprehensive, yet more central in its choice of topics than any outline of kings and their battles, and as much fun to read. A good deal of the fundamentals of how important things work, from fire flints to infectious diseases to steam engines, comes along with the chronicle.

WINDMILLS, BRIDGES, & OLD MACHINES: DISCOVERING OUR INDUSTRIAL PAST, by David Weitzman. Charles Scribner's Sons ($13.95). Keyed sectional drawings of an Adirondack iron furnace of 1849 (it stands a ruin near Tahawus, N.Y.) and a fine old Atlantic locomotive, a speedy 4–4–2, set the tone of this appealing book for junior-high archaeologists and up. The concise account of the evolutionary growth of industry, say for locomotives from *Tom Thumb* to *Big Boy*, offers not only perspective but also a start toward understanding the technology. The author takes some pains to give a very simple account of how steam engines and canal locks, boilers and smelters, and the structural trusses in bridges and roofs actually work. The simpler versions of the past are a more manageable starting place for the particular go of these devices than the larger and more complex examples of our time. It may seem commonplace to cross an unimpressive little bridge, but "you're a historian, an industrial archaeologist. You walk back a few yards, climb down a grassy embankment a ways, and...." Maybe you've found, like the author, a Whipple truss, modified for cast and wrought iron, the rods going every which way, the oldest iron bridge still in service in California.

The library too is a hunting ground for the search. Hidden in the books and magazines are plenty of old devices, and what people thought of them a century and more ago. "Old copies of *Scientific American* are a good place to start." Even better than the author's similar guide for adults, this small book is full of interest for mere readers and is better still for little teams of kids eager to get out into the field to some purpose.

FIREWORKS! PYROTECHNICS ON DISPLAY, by Norman D. Anderson and Walter R. Brown. Dodd, Mead & Com-

A carat or more.
Because you were never very good at fractions.



Your jeweler can show you many diamond solitaires of a carat or more. This necklace features a 1.14 carat diamond.

A carat or more–one in a million.

Every diamond is rare.
But of all diamonds found, a solitaire of a carat or more is only one in a million.

And, like love, becomes more precious with time.

A miracle among miracles.
Born from the earth. Reborn on a woman.

The extraordinary diamond of a carat or more.

Show the world you couldn't have made it without her.

A diamond is forever. De Beers

pany ($9.95). It is the family Grucci of Bellport, N.Y., that holds high the banner of American display fireworks. Here they are, with their long racks of wired mortars and the factory too, in photograph and text. The brief history of fireworks opens with some account of the deep past, back to the invention of gunpowder, and proceeds to a very simple account of the kinds of fireworks and how they work. The photographs are more detailed than the text, which settles for a few sentences on fuel and oxidant; the assemblies are better treated than the chemistry.

Three small Iowa towns that burned down from Fourth of July displays gave the impetus to the fireworks control laws; their tale is here, along with other disasters. Handel's *Music for the Royal Fireworks* (in 1749, the display set by the famous Ruggieris) was played to a paroxysm of unscheduled explosions by the light of a burning theater. Fireworks even in these safe and sane States still inflict some 11,000 injuries a year, mostly from oversize illegal fireworks in imprudent hands. Readers in the middle grades will learn a good deal both about the ironies of technology and about the rudiments of how the striking displays work. This book is not at all for budding pyrochemists, however careful; the how-to-do-it chapter is a neat one about taking photographs of that spectacle high in the sky, with a comment on distance measurement by the delay of the sound.

THE MAKING OF RETURN OF THE JEDI, edited by John Phillip Peecher. Ballantine Books ($3.50). Since the spring of this year almost 100 million people have seen the remarkable images made by imaginative engineering that convey the contemporary fairy tale of George Lucas. Out of the Monster Shop in San Rafael in northern California flow the creature costumes of sculptured and fitted plastics, the Pig Guards and the Squidheads, off to the EMI Elstree Studios. Industrial Light and Magic prepares its ingenious models and its paintings on glass. On the largest sound stage in Europe principal photography began in the first weeks of 1982, after elaborate set construction.

This well-informed chronological account of the 30 months from plan to preview centers on the sense of the people involved, with interviews with Lucas himself, with all the stars now so well known and a variety of specialists: the wizards of so many effects, stunts and sounds strange though real. Most of the story tells of telexes, urgent or bantering, of caterers, carpenters and little people recruited by the score to don the furry costumes of the Ewoks of the forest. One learns almost in passing about the "bluescreen" superposition scheme that created the speeder bikes that whiz through the forest, or about the radio-controlled pupils and eyelids of the grotesque tyrant Jabba the Hutt. Jabba was a jellylike slavering puppet large enough to contain three operators, one in each arm and a little one in the tail, each equipped with a video monitor to check on the effect outside. Two more operators worked those eyes from a distance.

Fantasy on such a large scale is exacting work; it is extraordinary to read of the care with which a videotaped study was made, complete with a miniature forest and cutout actors on hand-moved bikes, only as a guide to how the redwoods and the speeders would one day look in full scale but superposed. The actors had to mount stationary speeders in front of the bland bluescreen, of course seeing no trees at all, while they tried to imagine their feelings on a dizzying ride among the great tree trunks.

Pictured fantasy has power. Mark Hamill, who plays the last knight, holds the view that young people regard the film as real: "They don't want to hear about bluescreen techniques. To them, those are actual rocket bikes in an actual forest." This book is an easy-to-read account of the creation of a public myth, absorbing to any screen-struck young reader who wants nonetheless to part image from reality in a galaxy not at all far away.

Symbols and Stories

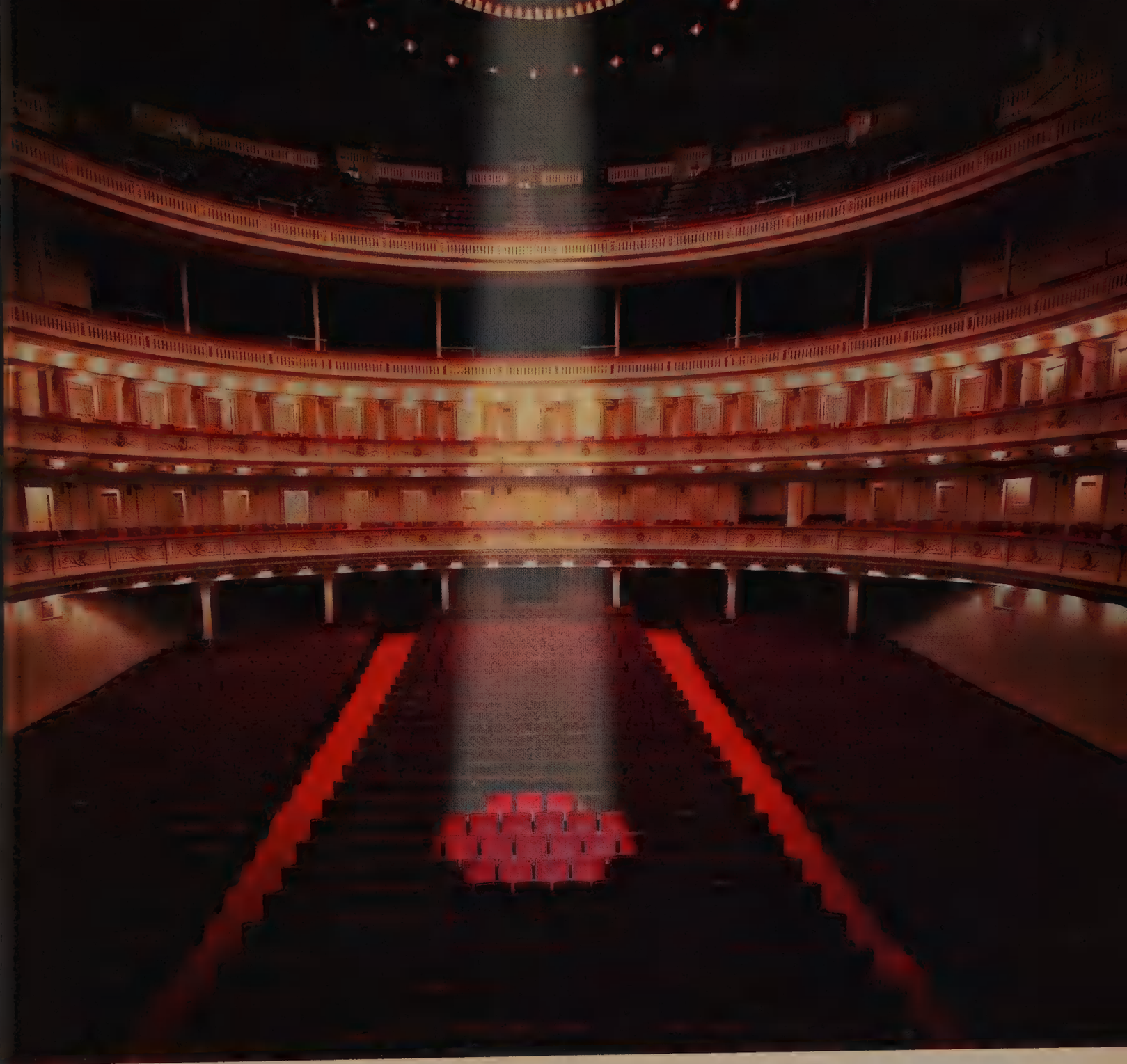
ANNO'S COUNTING HOUSE, by Mitsumasa Anno. Philomel Books, New York ($12.95). ANNO'S MYSTERIOUS MULTIPLYING JAR, by Masaichiro and Mitsumasa Anno. Philomel Books, New York ($10.95). These two books, one without words, one with a few lines of text, glow alike with the fine-grained paintings of the gifted Tokyo artist and traveler Mitsumasa Anno. In the second book his magic found assistance from his son.

There are really two counting houses, one timbered, one stone. They are drawn on the facing pages of the slender book, the facades removed, so that we see the rooms and their contents, including five boys and five girls, each with a number of attributes in common. The facades of the two houses are painted on the two sides of the leaf between every facing pair of cutaway houses. As you open the book anywhere, you see the colorful front of one house and the open interior of the other. The open section has no color, unless in one or another room there stands one of the colorfully dressed boys and girls. The 10 wander from house to house as the pages turn, carrying with them a picture of the Queen, a guitar or something else they love. Die-cut window openings through the facades expose more or less at random a view into one or another room on the page below. On this theme of 10 conserved children and their whimsical possessions there is much to be played out, up to the whole of sums and differences among 10. Anno makes it plain that his ingenious book is meant to lead children to find for themselves the beauty and pleasure in the world of numbers. (That royal portrait ends up in a closet.)

The multiplying jar is an actual setting of a task that is almost set by the nursery-rhyme traveler to St. Ives. Here a magical jar—very blue-flowered—holds a sea with 1 island. On the island are 2 countries, each with 3 mountains, every mountain with 4 walled kingdoms, and so on, until at last we find 9 boxes of 10 jars each. How many jars were there within? The answer comes by way of visual multiplication, dots drawn in arrays, up to two entire pages of dots for the 8! cupboards. There is not enough room for all the dots of the 9! boxes and the 10! jars, but the numbers are worked out and the idea of the factorial exclamation point is explained. This book repays early-grade skill at reading and at figures. Both volumes are arithmetic gems; no better introduction to the deep art of counting may ever have been published than those two houses shared among 10 cheerfully roaming children.

ORDERLY TANGLES: CLOVERLEAFS, GORDIAN KNOTS, AND REGULAR POLYLINKS, by Alan Holden. With photographs by Doug Kendall. Columbia University Press ($19.95). Three-space supports a marvelous variety of interlinked structures. The construction and properties of some of them are here exemplified and analyzed with elegance by a well-known writer on the pleasures of polyhedrons. The first half of this book of fresh recreational geometry, entirely without formulas, the eye always guided by crisply pictured models, treats of useful tangles. The first class of tangles is the cloverleaf highway intersection, realized so often and on such a large scale. Simple rules of traffic flow induce those looped connections; a practical degree of redundancy makes the motorist less likely to get dizzy as he shifts direction through many loops. The 12 one-way connections to link four lanes to each of the other three are in practice cut to eight.

Even more useful is the grand tangle we call woven cloth. Twill and tabby are here, constructed, like most of these structures, as well-made models formed of wood dowels instead of rope or yarn. We see also neat models of knots such as the square knot and the granny, not to mention the Turk's head and other baroque Gordian knots, those that cannot be untied without cutting, and we read something of their theory. A striking chapter treats cat's cradle, that ancient play of string-looped hands, as an example of unknotted rings; models, again in



Sony creates seventh row, center. Forever.

INTRODUCING THE SONY COMPACT DISC PLAYER.

You are looking at an invention so extraordinary that High Fidelity magazine has hailed it as "the most fundamental change in audio technology in more than eighty years."

A piece of audio equipment that permits you to hear something you've never heard before: perfection.

Sony is pleased to raise the curtain on the world's first digital audio compact disc player. The CDP-101.

THE END OF HI-INFIDELITY.

The CDP-101 uses a laser beam to play compact discs that hold far more musical information than conventional records ever could. The result: fidelity and depth of sound that no conventional turntable could ever deliver.

This highest of fidelity remains faithful too. Because the digital discs are read by laser beam, there is absolutely no physical wear. Nor is there any distortion from such annoyances as dust, fingerprints or even scratches. The laser simply sees through them.

Equally ingenious, an infrared remote control lets you select tracks without budging from your armchair.

While an ever-expanding library of compact discs lets you listen to your favorite artists as though you, and your armchair, were centered in the spotlight above.

Maybe most important, the CDP-101 is fully compatible with the conventional equipment you may own.

We suggest you hear the Sony Compact Disc Player soon. For a sound you can't believe, from the audio innovator you assuredly can.

SONY®
THE ONE AND ONLY



© 1983 Sony Corp. of America. Sony is a registered trademark of the Sony Corp. 1 Sony Drive, Park Ridge, New Jersey 07656.

wood, reveal the symmetry of the seven stations of the game with unexpected clarity. Finally, polylinks, the author's form of elegant chain, are defined, explored and illustrated in the rest of this highly original book.

MATHEMATICAL SNAPSHOTS, by H. Steinhaus. With a new preface by Morris Kline. Oxford University Press ($7.95). This member of the small set of mathematics brought to the amateur is offered in paperback for the first time, in the version first published in 1969. The same book, then a little smaller in its first appearance in English, was reviewed with enthusiasm by *Scientific American* in 1950.

A dozen swift, concise chapters, with a wealth of simple, clear diagrams or photographs, span a wide class of mathematical results, from Pythagoras to probability. The results are often surprising, occasionally profound and never dull. The key word is snapshot: in a few sentences of straightforward if tight argument and a clever sketch or two Professor Steinhaus makes his points, time and time again leading the reader to direct conviction and delight. Formal proof seldom occurs, no lines of algebra string across the pages, no long series of constructions appears. Open to readers equipped with high school algebra and geometry, the book is at once rich and disarmingly unpretentious. A fair share of the material will have meaning for mathematically inexperienced but interested readers.

To sample the hundreds of problems: how to share a cake with total fairness among any number of people; how to measure the great diagonal of a shoebox with a ruler too long to fit inside; how to trisect an angle to one part in a few hundred; how to list the Platonic solids without relying on the geometry of a plane (by dividing a sphere topologically); how to circumscribe Lake Michigan, or any other closed curve, within a square. This small classic of recreational mathematics is a find for anyone pleased with that witty domain who has not yet encountered these snapshots of what mathematicians really do and genuinely enjoy.

YEH-SHEN: A CINDERELLA STORY FROM CHINA, retold by Ai-Ling Louie. Illustrated by Ed Young. Philomel Books, New York ($10.95). Yeh-Shen, poor girl, did the heaviest chores for her jealous stepmother, there in the home of her late father, a chief of the cave people of southern China. The plot is well known: there is a magic benefactor, here no fairy godmother but a wonderful golden-eyed carp. Indeed, the poor carp was stabbed to death by the wicked stepmother, but its bones remain able to counsel the bright and lovely Yeh-Shen. Sure enough, there is a festival banquet. Yeh-Shen appears at it unexpectedly radiant in a gown of azure with a cloak of kingfisher feathers and a tiny pair of magic golden slippers. The tale unfolds without surprises or pumpkin coaches, the slipper lost as the unknown girl runs off in rags, the slipper reclaimed to fit the right tiny foot as the king finds his true love.

The luminous paintings glow across pages arranged to evoke an old Chinese book, although Ed Young's style in pastels and watercolor is richer than it is spare. The text is based on a Ch'ing edition of an encyclopedic work that retains the T'ang story by Tuan Ch'eng-Shih, more than 1,000 years old; two printed pages of the Chinese text are reproduced. Cinderella is found first written down in Europe in an Italian tale of the 17th century. The sense of the storyteller's unity across the world is strong here, in a very attractive book delightful on its own terms, even if you had never heard of Cinderella.

ZOUNDS! THE KIDS' GUIDE TO SOUND MAKING, by Frederick R. Newman. Illustrated by Elwood H. Smith. With a six-inch floppy stereo disc. Ran-



dom House, Inc. ($4.95). Once adept at the palate grind, the familiar pucker whistle, inhaled speech and the cheek thump—all of which are carefully described here as easy elements of spoken sound beyond everyday speech—any kid is on the way to evoking the night noises of the woods and the home, the fearsome roars and hisses of the jungle, the animal and mechanical noises of the farmyard, an entire band of musical instruments and the strange sounds imputed to space in the best-loved media. For each of these families of sound there is offered a small story, weaving the sounds into a more or less coherent plot. Just as in ordinary speech, the combinatoric power of a limited repertory (not, to be sure, just the basic four mentioned above) is the basis of elaboration. For the sound of a gently landing flying saucer try humming as you produce a pucker whistle, and let both glide down in pitch.

The book is written in a chatty, direct way; the illustrations are bold cartoons of a somewhat outlandish kind, as befits the buhs, ta-tas, ooo-wees and chiks that punctuate the text. The language is simple, although the oral tasks set are not easy for everyone to follow. They are perhaps easier for the aspiring young sound maker than for those nearby! This is an original and amusing guide to a diverting imitative art. The author is a lifelong artist in this domain; the accompanying recording bears cacophonous witness.

WHAT'S WHAT: A VISUAL GLOSSARY OF THE PHYSICAL WORLD, edited by Reginald Bragonier, Jr., and David Fisher. Ballantine Books ($9.95). THE OXFORD-DUDEN PICTORIAL GERMAN-ENGLISH DICTIONARY, edited by the Dudenredaktion and the German Section of the Oxford University Press Dictionary Department. Oxford University Press ($29.50). The scholars of Laputa, to preserve breath and minimize ambiguity, never spoke but carried with them their nouns, that is, an example of everything they might choose to converse about. Next come the big unabridged dictionaries of our youth, with thumbnail drawing after drawing to illustrate at least on paper the strange objects they define. Who has not looked up toughies, such as astragals and topgallants, in this kind of drawing?

These two reference books, one American and one Anglo-German, now aim to fill exactly such a role. The Ballantine book is a version of the Laputan backpacks, of course much more sensibly done through images on paper. Its hundreds of photographs show ice skates and the space shuttle, the tepee and playing cards, and the parts are named. Most of the pages are photographic images, so that the old dictionary thumbnail sketches are not visually invoked. One has the feeling that odd words are sought out and that the blander terms used in context to designate all the pieces of a toaster, say, do not add much knowledge.

The Oxford-Duden book consists almost entirely of crowded ink drawings, quite like those of the big old dictionaries. The pages are rich; in general each page shows a complex of items rather than a collection of single ones. One sees a hotel lobby or a cement factory, and keyed to numbers on each drawing are 50 names, both in English and in German, for key rack and bar man, for hammer mill and clinker cooler. The book was motivated by the belief that the visual invariants would help the task of technical translation; indeed, there is a version in English and Japanese.

The Oxford-Duden indexes about 28,000 terms, some 50 percent more than the Ballantine. The Ballantine, however, looks altogether more familiar to the Yankee eye, and it has a lighter touch; the Oxford-Duden never smiles. Either form of this remarkable mapping of the world is something kids would pore over and get much out of, although not always what was intended.



INTRODUCING THOROUGHLY LUXURIOUS 4-WHEEL DRIVE.

You're looking at what probably should be your family's next new vehicle. If you appreciate things like: lots of room, lots of comfort, lots of style, lots of fuel economy, lots of power and agility. The new leaner, meaner Jeep® Wagoneer is all of this and more. It's true, high-luxury. With a smooth car-like ride on the road, total comfort for 5 passengers and the ease of 4 doors. Plus Wagoneer gives you Shift-on-the-fly between 2- and 4-wheel drive as standard equipment. So with total convenience, you can both maximize your mileage and maximize your traction. The all-new Wagoneer. Leaner, meaner and thoroughly luxurious. At your Jeep dealers, who are proud to offer you the world's largest selection in 4-wheel drive. Buy or lease one today.

24*/32 EPA EST MPG EST HWY

THE ALL-NEW WAGONEER

WHEN IT COMES TO 4-WHEEL DRIVE ONE WORD SAYS IT ALL JEEP

Jeep is a registered trademark of Jeep Corporation. *Use figures for comparison. Your mileage may vary with speed, weather, trip length. Actual HWY and CA figures lower.

PROOF CALVERT DISTILLERS CO., N.Y.C.



Now that they're ready for a change of pace it's time to give them John Jameson.

Introduce them to John Jameson. They'll like the light, delicate taste. Luxurious and smooth as they would expect a premium whiskey to be. But with a distinctive character all its own. Step ahead of the crowd—give John Jameson, the world's largest selling Irish Whiskey.

Established 1845 SCIENTIFIC AMERICAN December 1983 Volume 249 Number 6

Modern Icebreakers

They have evolved rapidly because of economic pressures to maintain commerce in icebound waters. They feature unusual hull designs and propulsion systems, and most of them are currently built in Finland

by John D. Harbron

A substantial fraction of the world's seaborne commerce, particularly in the waters of North America and northern Europe, moves in the high latitudes where ice is an impediment for several months of the year. The nations engaged in such commerce deal with this impediment with a fleet of about 100 icebreakers, which require for their difficult task not only highly specialized construction but also unusual techniques of seamanship. It seems likely that the exploitation of resources such as oil in several regions in the high latitudes will intensify the demands on the icebreakers and their crews in the coming decades.

The distinctive features of an icebreaker are its shape, its strong bow, its machinery and its propellers. In shape the icebreaker is characteristically short in length and wide in beam, displaying therefore a stubby profile. The shape is accentuated in modern icebreakers built in Finland, whose shipyards have turned out some 60 percent of the world's fleets, by the piling up of the bridge and the crew's quarters above the deck. This design creates space in the hull for the sturdy main engines, the auxiliary engines and other machinery and the propeller shafts.

A look at an icebreaker in profile when the ship is out of the water would reveal another feature of the shape, namely that the draft is quite shallow at the bow, sloping to a greater depth amidships. This shape is made necessary by the fact that an icebreaker does much of its work not by driving head on into the ice but by riding up on it, so that the increasing weight soon causes the ice under the bow to break. The ship's unusual width contributes to the success of the stratagem by making the weight compact. The design also reduces the strains put on the hull by the fact that the ship is in effect repeatedly running aground.

Understandably a vessel doing this kind of work needs a bow that is thick, made typically of steel plate that is about five centimeters (two inches) thick and strongly braced with steel structural members. The hull is also reinforced with thick plate all along the waterline because that area will be often in contact with ice or struck by floes that have been broken loose by the action of the bow. At one time many icebreakers had a weakness in that the welds between the steel plates tended to corrode and fail. Workers at the foundry in the Rautaruukki primary steel mill in Finland developed a microalloyed steel in which the content of manganese, sulfur and silicon was limited compared with conventional steels. The welds in this plate have proved to be resistant to the low-temperature seawater in which icebreakers usually operate.

The propulsion machinery of an icebreaker faces severe demands arising from the extreme variations of load on the engines between the times when the ship is breaking ice and the times when it is backing off for another run or is operating at a forward speed in open water. To meet the heavier demands the engines are designed to reach full speed rapidly. In contrast, the engines of a conventional vessel work up slowly to full speed and could be damaged by having to develop maximum power rapidly. The icebreaker's engines must be robust, reliable and able to deliver full power over a range of revolutions. They also must be economical in fuel consumption, since the ship is often operating for extended periods in areas that are remote from fuel supplies.

In the days of steam the icebreaker had reciprocating steam engines. Now the diesel-electric system predominates. In these vessels diesel engines power electric generators, and the electricity runs the motors that turn the propellers. A few icebreakers with nuclear power have been built by the U.S.S.R., beginning in 1957 with the *Lenin*. The nuclear reactor in such a vessel is a source of heat for generating steam that runs turbines. Nuclear power is advantageous when an icebreaker is likely to do much of its work in areas remote from fueling stations, such as the high Arctic regions of the U.S., Canada and the U.S.S.R., but otherwise it cannot compete economically with diesel-electric systems.

An icebreaker may have as many as four propellers. If there are four, it is likely they will be "wing propellers," meaning that they are mounted on the sides of the ship, two near the stern and two near the bow. Bow propellers work well in first-year ice, that is, ice that has built up in a single winter and is not more than about 30 centimeters (12 inches) thick. (In some places, such as the Canadian Arctic, first-year ice is much more formidable, attaining a thickness of as much as two meters.) The bow propellers draw water from under the ice, weakening the support of the ice sheet and causing pieces of it to break off. Then the propellers wash water and ice away from the bow of the ship, facilitating forward progress. Reversed, they help the ship to move astern, thereby increasing its maneuverability. Bow propellers have proved to be useless in Arctic and Antarctic waters. The Arctic ice is too hard and the Antarctic ice is too thick to be affected by their action.

There is no single answer to the number of stern propellers an icebreaker



FINNISH ICEBREAKER *Urho* **opens a new channel through ice in the Gulf of Bothnia, followed by two cargo ships. The** *Urho* **was the first ship of a class of icebreakers, all of which have been built by the Wärtsilä shipyard in Helsinki. The ships are fairly large as icebreakers go: 104.6 meters (343 feet) long and 23.8 meters wide. Each of the ships has two propellers at the bow and two at the stern, with an output of 22,000 shaft horsepower. A diesel-electric system furnishes the propulsive power. Finland keeps the gulf open all winter.**

should have. The amount of propulsion they must deliver depends on the severity of the ice conditions where the ship is likely to operate. Several types of icebreaker with twin shafts, including the U.S.S.R.'s *Moskva* class in the 1960's and the Canadian Coast Guard's *R* class in the 1970's, have had a good deal of trouble with bent and broken propeller blades and fractured propeller shafts.

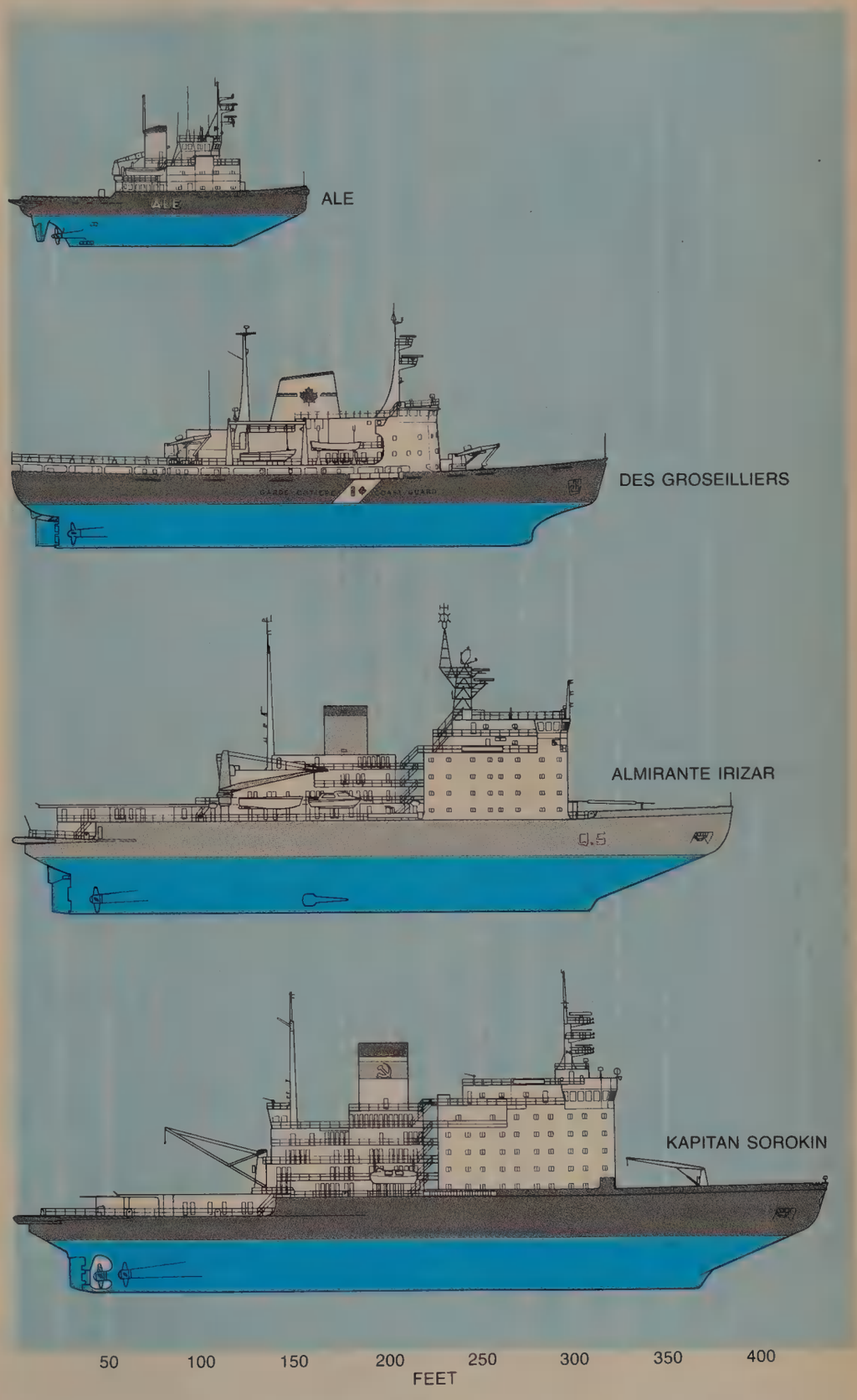
One answer to this problem has been to build icebreakers with three propellers, one in the conventional position at the middle of the stern and two on the wings. This configuration too has had trouble in the two icebreakers of the U.S. Coast Guard's *Polar* class. One reason apparently is that because of the dual role of these ships (to break ice in the Arctic and to serve as supply and research vessels in the Antarctic) they have too much shaft horsepower for the size of their hull. The result is a chronic vibration of the hull. Each vessel has 60,000 shaft horsepower (more than any non-nuclear icebreaker), 42,000 from the gas-turbine engine driving the main shaft for breaking ice and 18,000 from diesel-electric engines driving the two wing propellers for running in open water, as on the long trip to the Antarctic.

Since an icebreaker is highly specialized in construction and faces demanding sea and ice conditions, operating one calls for special types of seamanship. Examples are the two modes of operation in ice, known formally as the continuous mode and the ramming mode. In the continuous mode the ship moves steadily ahead at a speed of from three to five knots through ice that is as much as 1.5 meters (five feet) thick. The technique works if the ice is fairly level and offers only average resistance to the ship. Such conditions are normally encountered in several areas where icebreakers operate regularly, including the Gulf of Bothnia, the Gulf of Finland, parts of the Russian Arctic and the Canadian Arctic, the Great Lakes and the St. Lawrence Seaway.

In the ramming mode the ship rides up on the ice. When the ice breaks and the vessel is again level in the water, it is backed off several times its length and then driven forward to ride up on the ice sheet again. The technique has to be applied when the ice is notably thick, hard or irregular in surface form.

Yet another challenge to seamanship comes when the ship gets stuck in the ice, a not unusual occurrence even though the vessel is designed to break ice. (The *Polar Sea* of the U.S. Coast Guard was embarrassingly beset in the Chukchi Sea north of Alaska from February 18 to May 13, 1981. During much of that time the Coast Guard made the best of the situation by redesignating the ship as a floating research station under the label "Polar Sea Drift Project '81.") Having propellers at both ends of the ship can be helpful by increasing the vessel's maneuverability so that it is less likely to become icebound or has a better chance of breaking free. Another technique still in use on some older icebreakers is heeling, that is, causing the ship to rock from side to side by pumping water or fuel oil back and forth between heeling tanks on opposite sides of the craft. With trimming tanks fore and aft the ship can similarly be made to rock from bow to stern.

On more modern vessels the heeling



FOUR TYPES OF ICEBREAKER are portrayed at a common scale. The *Ale* class is designed for icebreaking assistance and towing duty in rivers, harbors and lakes in Temperate Zone waters such as the Great Lakes and the Baltic Sea. The *Des Groseilliers* is representative of the Canadian Coast Guard's *R* class, operating in the St. Lawrence Seaway and during the summer in the Arctic. The *Almirante Irizar,* built by Wärtsilä for Argentina, serves in the Antarctic both as an icebreaker and as a research ship. The *Kapitan Sorokin* is a member of a class of large icebreakers operated by the U.S.S.R. The sloping bow, heavily plated for repeated encounters with ice, enables an icebreaker to ride up on the ice, which breaks under the weight.

technique is being replaced by an air-bubbling system developed and patented by Wärtsilä, a Finnish shipbuilding firm that is the world's largest builder of icebreakers. The aim of the system is to reduce friction between the hull and the ice or snow. Compressed air is forced through a series of nozzles at intervals near the bottom of the ship. When the air rises along the hull in the form of bubbles, it creates a strong current of water and air that forms a lubricating layer between the hull and the ice. At low speeds (less than two knots) the system reduces friction to about half of what it would be otherwise. As the speed of the ship increases, the effectiveness of the air-bubbling system diminishes, but even at five knots the friction is reduced by about 15 percent.

Most people think of the ice covering the sea as being a fairly homogeneous sheet varying mainly in its thickness. Actually it is found in formations ranging from the impenetrable shelf ice of the Antarctic (some of it 100 years old) to the rough polar packs of the high Arctic and the winter-only ice of the Baltic Sea passages, the Great Lakes and the St. Lawrence Seaway. The people concerned with operating icebreakers have to take these many varieties of ice into account.

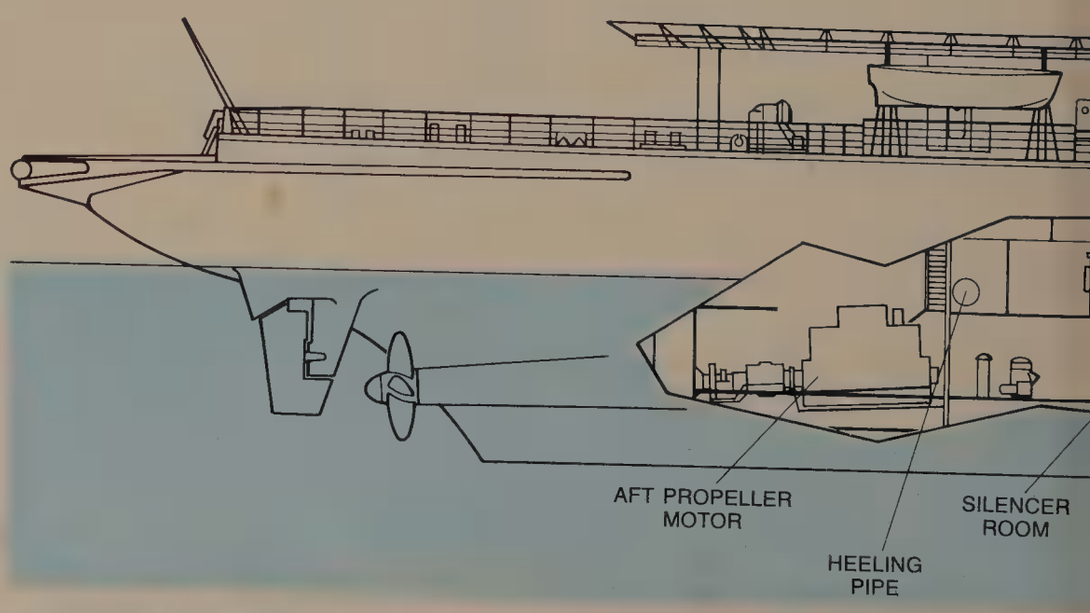
Until about 1966 the U.S., Canadian and Russian government agencies involved with shipping in the Arctic and the Antarctic employed only four rather broad classifications of ice formations: new, young, winter and polar. As the number of icebreakers increased and the means of determining ice conditions expanded to include surveys by aircraft and photographs from satellites, such generalizations proved to be inadequate for the crews that had to function successfully in icebreaking work. Now all the nations concerned with major shipping in areas where ice may be a factor maintain detailed and constantly changing definitions of ice formations. The definitions that follow are from Canadian sources, although the other nations recognize parallel categories.

The definitions begin with "open water," which is described in the *Sea Atlas of Canada* as "water in which the concentration of ice does not exceed one tenth of the total area." Another major formation is first-year ice, meaning ice that has persisted for less than a year. Second-year ice has survived one summer's melting and is building up again in the following winter. Multiyear ice has remained in place through two summers or more.

In addition to these kinds of ice there are other formations that can make an icebreaker's passage either difficult or impossible, depending on the extent of the ice structure. One is rafted ice, in which one sheet of ice has been overridden by another sheet or more than one. Another is hummocked ice, in which ice that has broken is forced upward by pressure and forms hillocks. A ridge is a wall of broken ice that has been formed by the collision of ice masses; its base can be as much as 15 meters below the surface. Finally, snow that has settled on an ice formation can considerably decrease an icebreaker's efficiency.

In general such regions as the Gulf of Bothnia, the Gulf of Finland, the Gulf of St. Lawrence, the Great Lakes and the northwest and northeast coasts of the U.S. have from 30 to 90 centimeters (about one foot to three feet) of ice between the middle of December and the middle of May. Severer conditions are faced by icebreakers in the Canadian Arctic and in the high-latitude waters of the U.S.S.R. and the Scandinavian countries, where efforts are made to maintain shipping for at least part of the ice season. Finland keeps the Gulf of Bothnia open throughout the year because of the need to move goods out of the country's many ports there.

A point that can be made about the Canadian Arctic also applies to other major ice areas. The Canadian Arctic extends some 3,000 miles from Baffin Island to the North Slope of Alaska and about 2,000 miles northward from Canada's Arctic coastline to the North Pole, with major passages and islands in between. The region has an archipelago of diverse ice-covered areas. Each area has its own historic record of ice, wind, sea-condition and other climatologic factors, and any one of them can change



BELOWDECKS MACHINERY of the *Urho* and other icebreakers of its class takes up almost all the space in the hull. It is largely to make room for this equipment that the bridge and the facilities for the crew are all above the deck on many icebreakers, giving them a top-heavy

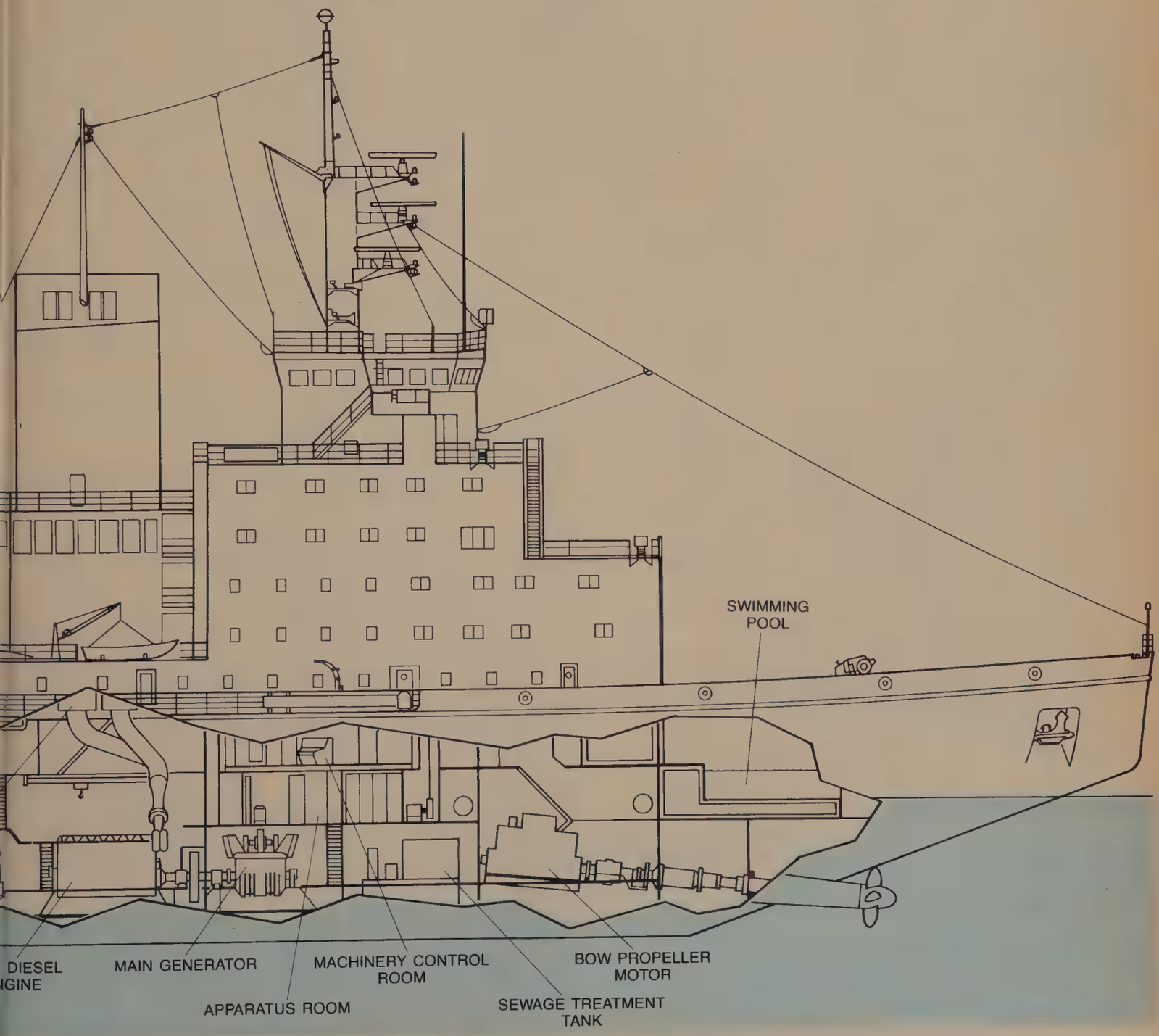
unpredictably in a particular season of icebreaker operations. T. C. Pullen, a retired captain in the Canadian Navy and a specialist on icebreakers, has summarized the situation as follows: "We encounter heavy ice where intelligence has reported ice-free conditions, providing yet another lesson in what Arctic marine operations are like, namely that they are not necessarily what we should assume to be the case."

The variations in the ice give rise to variations in the design of icebreaking vessels. Again the Canadian standards provide a guide. A few years ago the Canadian government published its Canadian Arctic Shipping Pollution Prevention Regulations, which combine a concern for the environmental safety of the Arctic as increasingly larger ships are active in it and an effort to specify the capability of Arctic ships. The regulations define icebreakers and their cousins, ships strengthened for ice, in a range from Arctic Class 1 to Arctic Class 10.

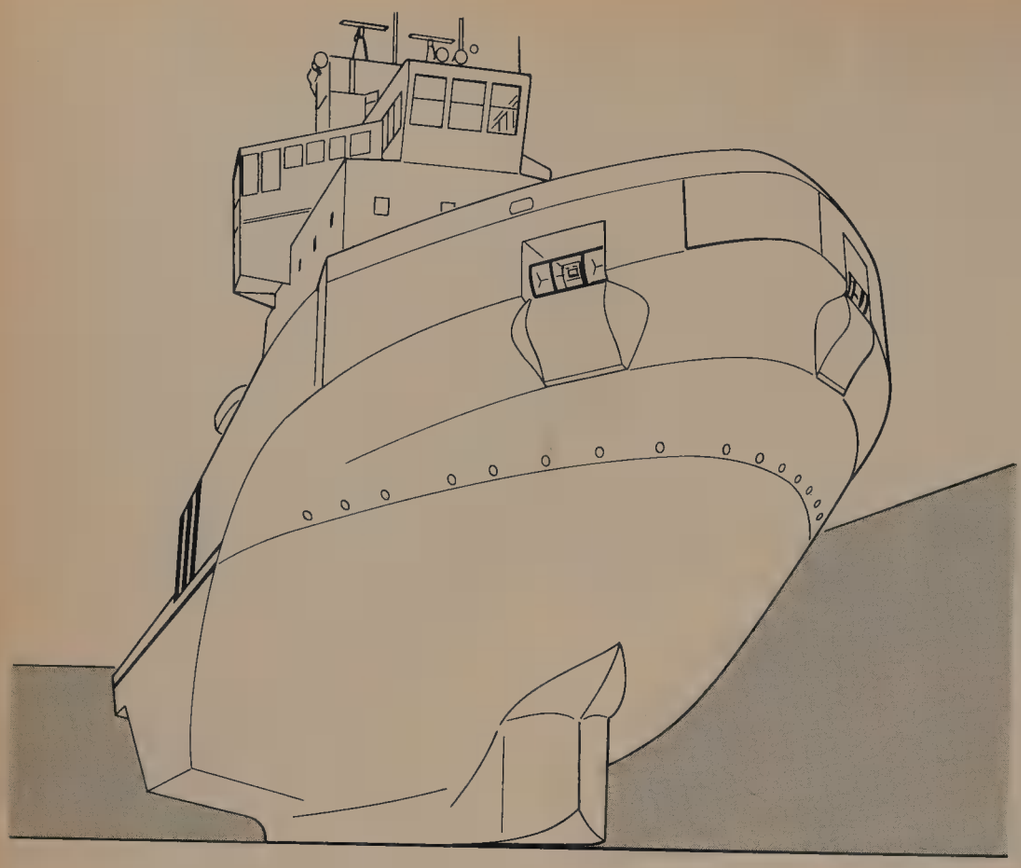
A few examples will indicate the gradient embodied in the regulations. A vessel of Arctic Class 3 can maintain a speed of three knots through ice .9 meter (three feet) thick. A vessel of Class 7 can maintain a speed of three knots in the continuous mode through ice 2.1 meters (seven feet) thick.

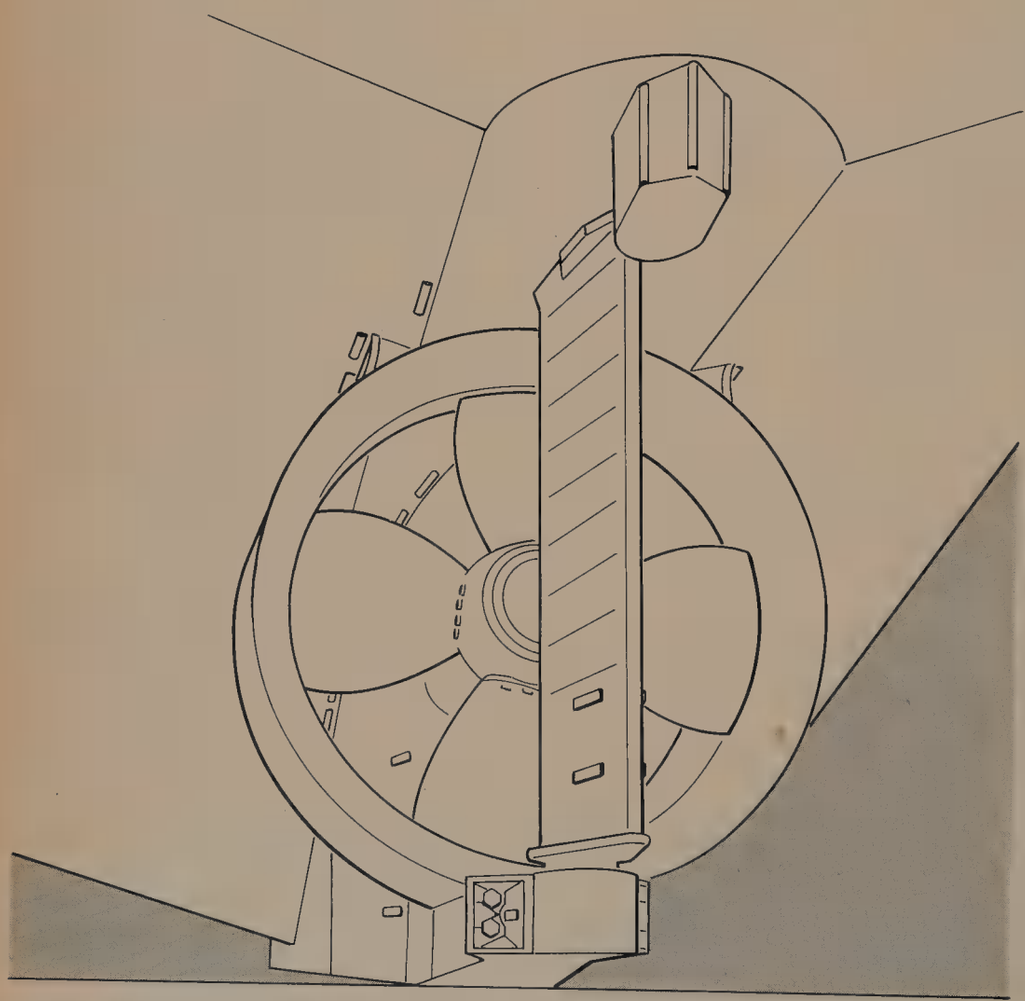
The icebreakers of the Baltic type built mainly in Finland and the U.S.S.R. to work in regions with one-year ice can be identified under the regulations as vessels of Class 3 or Class 4. They would not meet the more demanding requirements for a large icebreaker operating throughout the year in the Canadian Arctic. Vessels of this kind are in Arctic Class 8. Canada plans to build its first icebreaker of the class but has not yet begun construction.

The most powerful icebreakers today are the nuclear-powered Russian sister ships *Artika* (recently renamed *Leonid Brezhnev*) and *Sibir,* which can be estimated as in the range from Class 5 to Class 7 under the Canadian regulations. Each is a vessel of 35,000 tons' displacement (equivalent to the weight of the ship and everything that it is carrying) with 75,000 shaft horsepower. Another nuclear icebreaker, the *Rossiya* (40,000 tons and an estimated 100,000 shaft horsepower), is now under construction



appearance. The heeling pipes carry water back and forth between port and starboard tanks when it is necessary to rock the ship from side to side to avoid becoming stuck in ice. The role of the bow propellers is to draw water from under the ice ahead of the ship, weakening the support of the ice sheet and making it easier to break. Such propellers work best in ice that is no more than about a foot thick.





SPECIAL FEATURES of the *Canmar Kigoriak*, a Canadian-built icebreaker employed by Dome Petroleum Limited mainly to protect oil rigs in the Beaufort Sea, include a spoon-shaped bow, a special reamer and a single propeller surrounded by a protective structure. The bow design is meant to break thin ice with a minimum loss of energy. The reamer digs a short distance into the ice, providing a fulcrum that makes turning easier. The purpose of the propeller arrangement is to avoid the milling of broken ice that floats toward the stern; the milling action reduces the efficiency of the propeller. The *Kigoriak* also has an unusually large rudder.

for the U.S.S.R. at the Wärtsilä shipyard in Helsinki.

The *Rossiya* will emerge from the shipyard without engines. It will then be towed to Leningrad for the installation of the nuclear power plant. The Russians have not revealed the particulars of these marine nuclear power plants and do not allow foreigners to visit their nuclear icebreakers.

The Antarctic icebreakers of the U.S., Argentina, Japan and West Germany are multipurpose vessels compared with the ones that operate in the Northern Hemisphere. The reason is that they must not only deal with ice but also travel long distances in open water to get to the Antarctic, where they function additionally as observation, research and supply ships. Accordingly they are fitted out with extra cabin space for the scientific workers, landing pads for short-range helicopters that make reconnaissance flights over the ice, heavy deck cranes, cargo elevators and special holds for the supplies they carry to their nation's bases in the Antarctic.

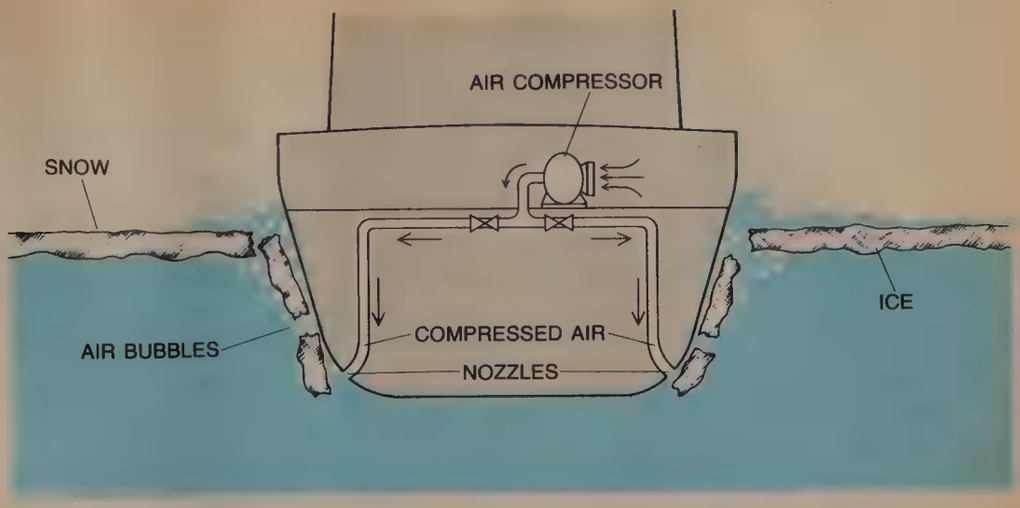
An example of the type is the Argentine *Almirante Irizar,* built by Wärtsilä and commissioned in 1979. One of its special features is a helicopter "touch down" system made in Canada. It is designed to cope with the fact that a shallow-draft, stubby icebreaker can roll as much as 40 degrees in open water when the seas are heavy. A cable is lowered from the helicopter to a locking mechanism on the deck, which slowly draws the aircraft down to the level of the deck. By means of a track set in the deck the helicopter is then towed to its hangar. The vessel also has two 16-ton hydraulic cranes for transferring cargo to the shore, an electric towing winch fitted with tension controls (keeping the cable taut at all times when it is loaded) and capable of exerting a maximum pull of 60 tons, and an air-bubbling system.

The U.S.S.R. is the major operator of icebreakers and the only nation with nuclear-powered icebreakers in regular service. It is also the only country operating significant numbers of merchant ships with a limited capability for icebreaking. (Many countries operate merchant vessels with hulls strengthened for ice.) Among the Russian vessels are a number of multipurpose container ships with icebreaking capability (at about the level of Arctic Class 2) that can proceed on their own in moderate ice conditions if an icebreaker is not available. Typically they are ships of 18,000 deadweight tons (the total carried load), shallow draft and an air-bubbling system. They can penetrate the large Siberian rivers flowing into the Arctic seas.

Russian icebreakers operate across the immense length of the northern sea route and also, during the ice season, in

such peripheral bodies of water as the Gulf of Finland, the Baltic Sea, the Black Sea, the Caspian Sea and the Yenisei, Ob and Lena rivers. The most impressive icebreaker-convoy operations are in the Russian western Arctic, where fleets of commercial vessels are moving in increasing numbers from Murmansk through the Barents Sea and the Kara Sea to the Yenisei River estuary, where they continue upstream to the interior port of Dudinka that serves the timber- and nickel-exporting Siberian city of Norilsk.

These integrated fleets function in the following way. Nuclear and diesel-electric icebreakers of the *Brezhnev, Kapitan Sorokin* and *Ermak* classes escort the container ships from Murmansk to the Yenisei estuary. The convoy may also include some European cargo ships, mainly British and Scandinavian vessels cleared under Russian ship classifications for navigation behind an icebreaker. At the Yenisei the nuclear icebreakers and the bigger conventional ones stand off the mouth of the river as the shallow-draft icebreakers of the *Kapitan Sorokin* class continue on into the estuary to the point inland where the convoy is taken over by river icebreakers of the *Kapitan Chechkin* class. They are vessels of 4,490 shaft horsepower and extremely shallow draft (3.2 meters, or 10.5 feet). If the ice conditions are too severe for the river icebreakers and the convoy to proceed to Dudinka, the cargoes are unloaded onto "polar utility vehicles," that is, hovercraft that skim over the ice on a cushion of air. The return passage down the Yenisei and the escort back to Murmansk by the heavier icebreakers complete the journey.



AIR-BUBBLING SYSTEM was designed by Wärtsilä as an alternative to the heeling technique for preventing an icebreaker from getting stuck in the ice. Compressed air forced through nozzles along the bottom of the ship rises along the hull in the form of bubbles. The result is a strong current of water and air that creates a lubricating layer between the hull and the ice.

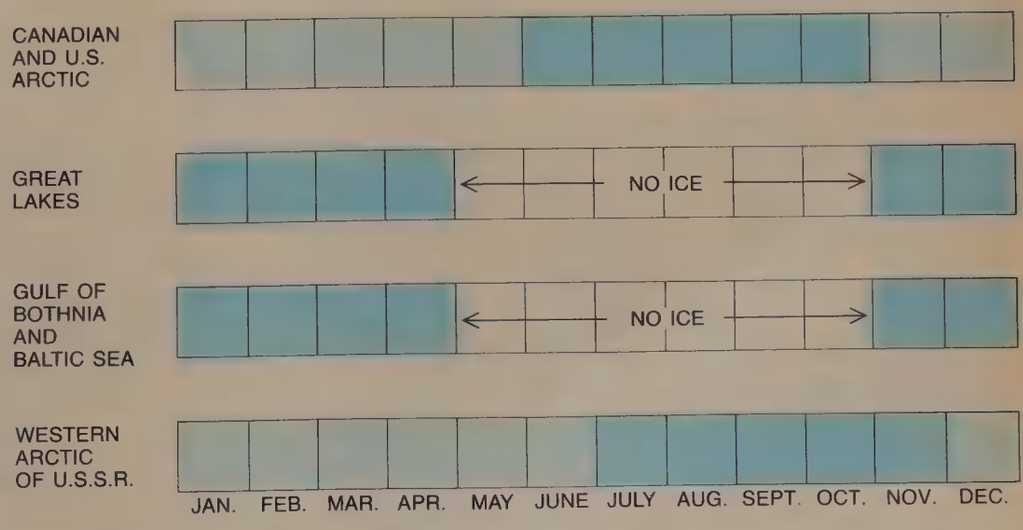
It may be that the characteristics of the icebreaker of the future are being developed by a Canadian vessel recently put in service, the *Canmar Kigoriak* of Dome Petroleum Limited. (The first part of its name is an abbreviation for Canadian Marine Drilling Limited, Dome's shipping subsidiary; the second part is the Inuit name for the northern lights.) This Canadian-built vessel serves mainly in the Beaufort Sea to protect Dome's offshore drilling rigs against damage by ice and to assist the firm's high-Arctic fleet, which includes ice-strengthened drilling ships, tugs and self-propelled barges.

Among the distinctive features the *Canmar Kigoriak* has for coping with the severe ice conditions of Canada's western Arctic is a single propeller, which contrasts with the twin or triple screws of all other Canadian icebreakers. Its designers see this arrangement as a means of avoiding what they call "one of the main perils of twin- and triple-screw icebreakers, the constant milling of ice with the two outboard screws, which reduces propeller efficiency." The propeller is surrounded by a large protective structure that deflects large pieces of ice moving aft and harmlessly mills small ones. The ship also has bow and stern thrusters and an outsize rudder, which increases its maneuverability.

Another feature of the ship is a hull form that distinguishes it from all other icebreakers. Its spoon-shaped bow is meant to break thin ice with a minimum loss of energy. A special reamer is fitted where the bow meets the midsection; it is designed to dig a short distance into the ice, thereby providing a fulcrum that improves the ship's ability to turn. The vessel's capabilities are indicated by the owner's statement that so far it has penetrated first-year ice as thick as 1.73 meters, multiyear ice as thick as 10 meters and first-year ridges grounded in 13 meters of water.

The *Canmar Kigoriak* is also a prototype for a giant ship that Dome may build in the 1990's, an icebreaker-tanker of 200,000 deadweight tons. The plan is based on Dome's estimate that enough offshore oil will be found in the Beaufort Sea to justify marine transporters of this size to move it out. Only ships of such size and enormous power will be equal to the task of operating year-round in the Northwest Passage.

If Canada's Arctic Class 8 icebreaker is built, it will be the largest non-nuclear icebreaker in the world. The *Polar 8,* which is its drawing-board name, would be 194 meters long, 32.2 meters wide and 13 meters deep. At 37,000 displacement tons and 100,000 shaft horsepower, the vessel would represent a substantial change in the standard formula for icebreaker design that relates displacement tonnage to shaft horsepower. In non-nuclear icebreakers a characteristic ratio is 15,000 displacement tons to 20,000 shaft horsepower, as in Wärtsilä's widely employed *Kapitan Sorokin* class now in service in Finland, the U.S.S.R. and Argentina. The *Polar 8* would be the first all-weather icebreaker capable of working in Arctic ice conditions that are too severe for the U.S. and Canadian icebreakers in service today.



SEASON OF OPERATION in the main icebreaking areas is charted. The months when ice is present are indicated in color; the dark color shows the months of icebreaker operation.

Io

The solar system's most active volcanic world is a moon of Jupiter whose geyserlike eruptions and massive lava flows yield evidence of exotic volcanic fluids. Their source of energy is gravitational

by Torrence V. Johnson and Laurence A. Soderblom

In the modern era of planetary exploration Io, a moon of Jupiter discovered by Galileo almost 400 years ago, has emerged as one of the strangest worlds in the solar system. First in the 1960's astronomers discovered that Io somehow modulates the intense bursts of radio waves emitted by Jupiter. In addition sensitive instruments trained on Io from the earth confirmed that the object is as reflective as fresh snow and yet its color is yellow orange. This raised perplexing questions about the composition of Io's surface. Then in the 1970's it was discovered that great clouds of ions and neutral atoms are being injected by Io into the Jovian magnetosphere, the region in which electrically charged particles are trapped by the magnetic field of Jupiter. Finally spacecraft entered the Jovian satellite system. The flybys of *Voyager 1* and *Voyager 2* in 1979 yielded the first close look at Io.

The Io we know today seems even stranger than before. Heated by tidal forces exerted by Jupiter, it appears to be the most volcanically active body in the solar system. It is clearly more active than the earth. Great volcanic geysers on Io send eruptive plumes to heights of hundreds of kilometers, and matter escaping from the tenuous volcanic atmosphere of Io by processes that are not yet well understood does indeed dominate the Jovian magnetosphere, thereby controlling the rate at which Jupiter radiates energy and probably affecting phenomena as diverse as auroras on Jupiter and radio bursts from the planet and its satellites. The discoveries made by the Voyager spacecraft have raised a host of new questions; hence this article cannot be definitive. It is instead a report on the continuing investigation of a remarkable celestial object.

Io is one of Jupiter's "regular" satellites: the ones that travel in essentially circular orbits more or less in the same plane as Jupiter's equator and so are thought to have formed around Jupiter much as the planets formed around the sun. At an orbital distance of 350,000 kilometers (about the same as the distance of the earth's moon from the earth), it is the innermost of the four largest Jovian satellites, which were discovered by Galileo. Io and Europa, the next outward of the four, are about the size of the earth's moon, and both have a rocklike density: respectively 3.5 grams per cubic centimeter and three grams per cubic centimeter. Ganymede and Callisto, the outermost of the four, are larger (they are about the size of Mercury), and they have a significantly lower density, about two grams per cubic centimeter. Apparently they consist half of water ice and half of rock, mostly silicates. The density variation among the four bodies has often been called a smooth trend, but in truth it is more of a bimodal distribution. The inferred mass of the silicates in Ganymede or Callisto is approximately the same as the total mass of Io or Europa; thus the four can be viewed as being similar except that the inner two are missing their share of volatiles, principally water. (Europa is known to be covered by ice, but its bulk density indicates that this covering is a veneer whose thickness is less than one 20th Europa's radius.)

The density variation is usually attributed to the effect Jupiter had on the early conditions in its vicinity. Even today Jupiter radiates at infrared wavelengths about twice as much energy as it receives from the sun; it has clearly been cooling slowly from an early luminous state. Hence models of Jupiter's early history (based on physics much like that of low-mass stars) suggest that the early Jupiter was an important local energy source for several hundred million years. Models developed by James B. Pollack of the Ames Research Center of the National Aeronautics and Space Administration and his colleagues suggest that over a period on the order of 100 million years the temperature and the pressure at Io's current orbital position were dominated by Jupiter, not by the sun. In fact, a body at Io's orbit in the first million years of that period would have received more energy from Jupiter than the earth receives now from the sun.

It is likely, therefore, that the composition of the regular moons of Jupiter was affected by their relative proximity to the planet. In this view Io and Europa are rocky bodies because conditions near Jupiter were too hot to allow those moons as they were forming to retain any substantial quantity of water. In contrast, Ganymede and Callisto, which did retain water, became mixtures of ice and rock and grew larger than Io and Europa. It is notable that the densities of Ganymede and Callisto are about what one would expect if a gas with the composition of the sun were cooled to the temperature at which ice condenses.

This simple view cannot be the entire story. For one thing the bodies in the solar system were intensely cratered by planetesimals in a period called the heavy bombardment, which ended about four billion years ago. Planetesimals falling into Jupiter's gravitational

IO'S SOUTHERN HEMISPHERE was photographed in 1979 by a wide-angle camera on the spacecraft *Voyager 1*. The field of view encompasses about 2.5 million square kilometers, or 6 percent of Io's surface; one centimeter represents about 50 kilometers. Several classes of features are seen. Calderas (volcanic-collapse craters) are the most numerous. Some are surrounded by diffuse surface markings, which probably represent the painting of the surface by sulfurous geysers. Others are surrounded by what are probably eruptive flows. Still others appear to be filled with eruptive material; the lozenge-shaped caldera with red and black patches in it is an example. White splotches dotting the surface are thought to be laid down by mixtures of sulfur dioxide gas and frost erupting along scarps and fractures. A mountain at the lower left is taken to be an outcrop of the silicate crust of Io and is estimated to be nearly 10 kilometers high. The image has been processed by a computer of the U.S. Geological Survey so that one views the various features as if one were hovering directly above the landscape.





PELE was the first volcano discovered on Io; it is also the largest geyserlike eruption observed there so far. The plume of the geyser is visible above the limb of the moon; it ascends to a height of about 300 kilometers and has deposited on the surface a concentric set of yellow and brown rings. The outermost ring averages 1,400 kilometers in diameter. At the center of the deposit is the source of the geyser: a complex of hills with a central valley. Markings and flows are evidence of past eruptions. This view of Pele was made by *Voyager 1;* it is a mosaic produced by a technique developed by Alfred S. McEwen of the Geological Survey in which high-resolution images contribute spatial detail and low-resolution images contribute color data. When *Voyager 2* arrived four months after *Voyager 1,* Pele was inactive.

field would have been greatly accelerated; hence Io and Europa may have received a more thorough battering than the Jovian satellites farther out. In addition the intense volcanic activity on Io has undoubtedly altered Io's content of volatiles. Still, it seems that Io, and more generally the regular moons of Jupiter, resulted from processes similar to those that gave rise to the planets. Jupiter and its moons are thus a miniature solar system in a deeper sense than the merely geometric one.

The surface Io revealed in Voyager images is dominated by volcanic activity to an extent that was almost totally unexpected. The reason is as follows. The normal influences on the thermal evolution (and so on the volcanism) of a planetary body are thought to be, first, the heating of the body by the release of gravitational potential energy in its accretion and also by the decay of short-lived radioactive isotopes such as aluminum 26; next, the gradual buildup of heat due to the decay of long-lived radioactive isotopes, mainly of uranium, thorium and potassium, and finally, a gradual cooling as the radioactive heat sources are depleted and heat is lost through the surface of the body by convection and conduction. It is a sequence favoring volcanic activity in large solid bodies, which have a greater ratio of volume to surface area than small ones. Hence the earth is fairly active today, whereas the moon's volcanic fires were banked long ago. (The ones on Mars are probably active today but are much less intense than the earth's.) Io, which has the size and the mass of the earth's moon, would be expected to exhibit no more than signs of ancient volcanism.

This expectation changed dramatically in March, 1979, just days before *Voyager 1* reached the Jovian system, when Stanton J. Peale of the University of California at Santa Barbara and Ray T. Reynolds and Patrick M. Cassen at the Ames Research Center published their analysis of tidal forces as a heating process for Io. Io and Europa are locked into one of the gravitational resonances first studied early in the 19th century by Pierre Simon de Laplace. In each such resonance the gravitational interaction of two moons adjusts their orbital distances until their orbital periods are common multiples. (The orbital period of Europa is twice that of Io.) One effect of the Io-Europa resonance is that Europa continually forces some eccentricity, or deviation from circularity, into Io's orbit. As a result the sub-Jupiter point on Io (the point on the body closest to Jupiter) oscillates, producing a tidal bulge that migrates back and forth on Io and varies in amplitude, heating

the body by friction. Peale and his colleagues calculated that the amount of heating could be surprisingly large, and so they predicted that significant volcanic features might be present.

As *Voyager 1* neared Io it became clear that Peale and his colleagues were right. In images of ever increasing resolution no traces of large-scale impact craters were found. Instead the surface was covered with varicolored red, orange and yellow splotches, which soon resolved themselves into features resembling terrestrial and Martian lava flows and calderas (volcanic-collapse craters). It was obvious that Io has a geologically young, active surface. A few days after the *Voyager 1* encounter a further analysis of the images by the Voyager navigation and imaging-science teams revealed active volcanic eruptions on Io. Since then the nature of the volcanic activity revealed in the Voyager images has been intensively studied, and a preliminary picture of the kinetics and thermodynamics of Io's volcanism has begun to emerge.

Many of the characteristics of Io's volcanoes stem from the exotic chemistry of its surface. Well before *Voyager 1* the absence of certain features in Io's infrared spectrum (specifically absorption lines due to water ice or frost) had shown that Io's surface is extremely dry, and details of its spectrum at visible wavelengths had led to proposals that sulfur in some form is there. Spectral-emission lines due to sulfur ions trapped in Jupiter's magnetosphere were also suggestive. Then a variety of instruments on *Voyager 1* and *Voyager 2* showed that large quantities of sulfur and oxygen ions are spread throughout the magnetosphere but are concentrated near Io's orbit. Ultraviolet emissions from such ions proved to come from a torus around Jupiter centered on Io's orbit. Finally, members of the Voyager infrared-spectrometer team discovered infrared absorption features due to gaseous sulfur dioxide over one of Io's volcanoes. This made it clear that sulfur and sulfur compounds are central to Io's volcanic activity. Spurred on by the Voyager findings, two research groups, one at the Jet Propulsion Laboratory of the California Institute of Technology and one at the University of Hawaii, identified a prominent feature in Io's infrared spectrum near a wavelength of 4.1 micrometers as arising from sulfur dioxide frost.

Why is sulfur so prominent on Io? In the universe sulfur is fairly common: it is only about 40 times less abundant than oxygen. In the crust of the earth it is scarce, but then terrestrial sulfur is thought to be mostly hidden in the core in the form of iron sulfide. On the moon too sulfur is scarce. There the dearth of sulfur is part of a larger problem: the moon's low content of volatiles, which remains one of the greatest difficulties facing the effort to explain how the moon originated. In meteorites, particularly primitive ones, sulfur is abundant, and on the surface of Mars data collected by the Viking landers reveal the abundant presence of sulfur in one form or another.

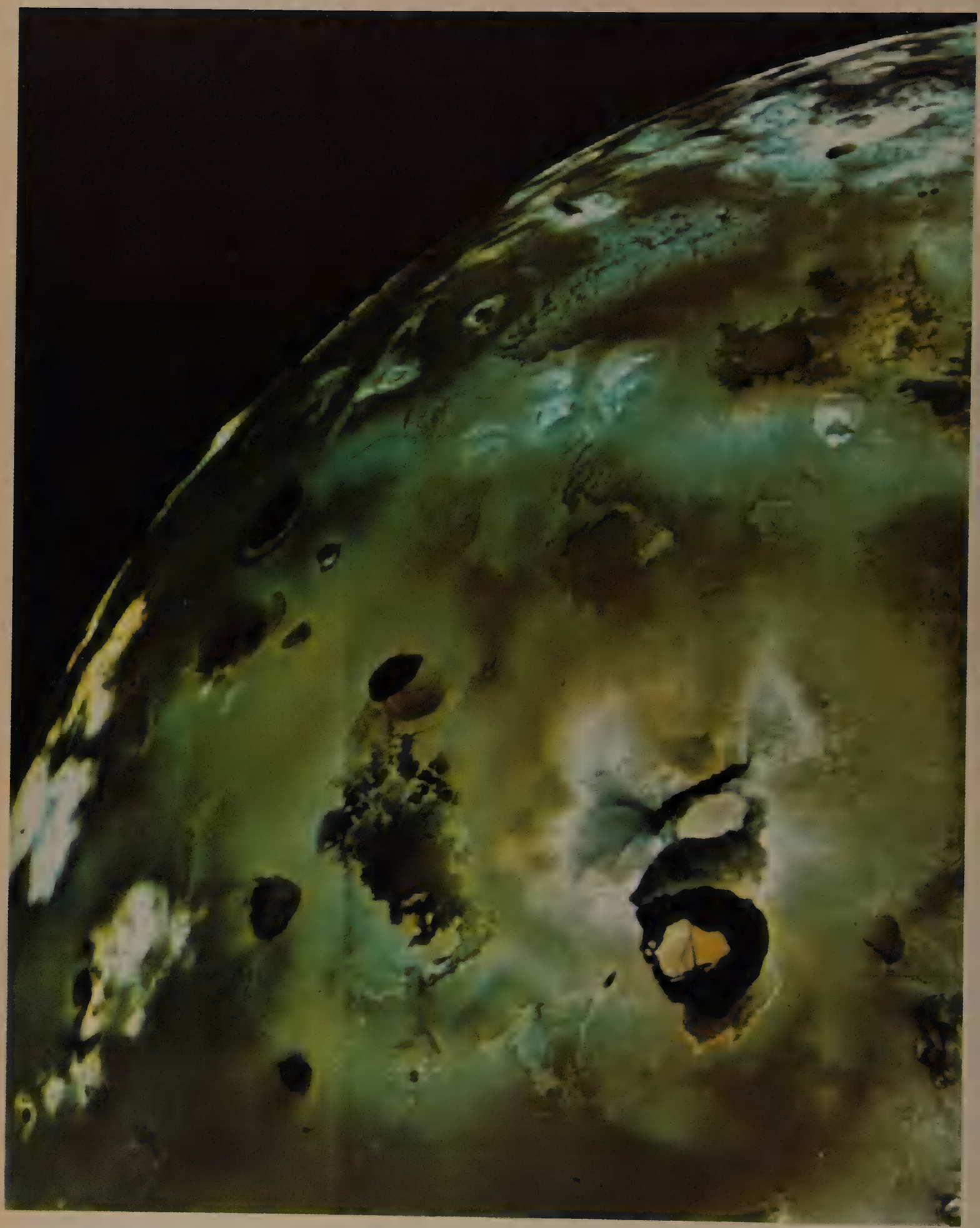
Most models of how the moons of Jupiter formed suggest that initially Io should have been richer in volatiles than the earth's moon. It must then be presumed that intense tidal heating over the history of the solar system has driven off all the light volatiles on Io, including water, so that sulfur and its various compounds are the only major volatiles left. A competing possibility is that conditions around the incipient Io were hot enough to prevent the inclusion of water but not too hot to prevent the inclusion of sulfur. The nature of Amalthea, the next moon inward from Io, may decide the issue: if the incipient Io was too hot for water, Amalthea should have been even hotter (assuming it formed in its current place with respect to the other Jovian moons), and so it should consist of very refractory matter indeed. It has been suggested, however, on the basis of limited spectral data that the surface of Amalthea is similar to primitive, carbonaceous meteorites, which implies a significant volatile content.

Several types of volcanic activity on Io must be distinguished; all are probably influenced by the chemistry of sulfur. One type, eruptive plumes, accounts for the most dramatic volcanic activity observed on Io so far. Consider Pele, which was active during the *Voyager 1* flyby and was the plume discovered first. It rose about 300 kilometers and distributed its ejecta in an umbrella-shaped pattern about 1,400 kilometers in diameter. The umbrella shape, which typifies many of the plumes, suggests that their ejecta travel ballistically. On the other hand, a set of images of Pele's plume records a swirling structure. The swirls suggest turbulence, and hence that the visible particles are interacting with some invisible propulsive gas. (The Voyager images show primarily solid particles of micrometer size or smaller; the associated gases would not be seen directly.) Computer models of plumes formed by ballistic particle ejection under Io's gravity match the general characteristics of Io's symmetrical plumes quite well.

Any model of the mechanisms responsible for the plumes must explain the high exit velocities derived from the ballistic analysis, which range between 500 and 1,000 meters per second. Bradford A. Smith of the University of Arizona and his colleagues Eugene M. Shoemaker and Susan W. Kieffer of the U.S. Geological Survey and Allan F Cook II of the Center for Astrophysics of the Harvard College Observatory and the Smithsonian Astrophysical Observatory have proposed a geyserlike model, taking as their starting point the inference that the great amount of heat being pumped into Io by tides must result in subsurface temperatures that allow liquid sulfur dioxide and molten sulfur to be in contact at relatively shallow depths. Accordingly Kieffer has made an intensive study of the thermodynamics of plume eruptions for a wide



PROMETHEUS represents a class of volcanic plumes on Io that are smaller, cooler and longer-lived than the ones resembling Pele. These three views of Prometheus were made by *Voyager 1*. The top view shows the plume in profile; it is 100 kilometers high and 300 kilometers across. The other two views show the plume as the spacecraft passed overhead. The plume is dark against the surface. Nevertheless, it has laid down a ring of bright material. The brightness probably indicates a surface accumulation of sulfur dioxide frost.



LOKI PLUMES emerge from the ends of a linear black fissure some 200 kilometers long. Each has produced a bright fan and a larger, darker deposit, as if the two plumes mixed the eruptive styles of Prometheus and Pele. Below the Loki fissure is a *D*-shaped black patch that may be a lava lake, although it remains to be established whether the lava is sulfur or silicate. Measurements of Io's infrared radiation made by the Voyager spacecraft indicate that most of the surface of the lake has a temperature of about 300 degrees Kelvin, which is 170 degrees higher than the temperature of the surrounding terrain. The image is a *Voyager 1* mosaic processed by McEwen's technique.

variety of conditions involving these substances. At the low-energy end of a range of possible eruptions is one that starts when liquid sulfur dioxide comes in contact with hot liquid sulfur. The liquid sulfur dioxide begins to boil, so that a mixture of liquid and gas begins to expand up some conduit toward the surface. As the expansion continues, the production of vapor increases. Then the sulfur dioxide attains its triple point: the temperature at which ice, liquid and gas coexist. At this juncture, still at some distance below Io's surface, the remaining liquid freezes, and the further expansion of sulfur dioxide is marked by the condensation of sulfur dioxide vapor into a snow of solid particles. An alternative higher in energy starts with superheated sulfur dioxide vapor. Here the expansion toward the surface is accompanied first by the condensation of liquid and then by the freezing and snow.

In each case the mixture of sulfur dioxide phases is accelerated inside its conduit until the flow reaches the velocity of sound (in that medium). At that point its velocity (a few hundred meters per second) can no longer increase. Nearer the surface the conduit may widen. Moreover, on reaching the surface the flow can expand into the tenuous atmosphere of Io. Hence the flow can attain an impressive velocity just above the vent. By calculating the amount of energy available Kieffer has estimated that exit velocities of more than 1,000 meters per second are possible. She notes, however, that a large range of subsurface conditions can account for the Voyager observations, so that it is not yet possible to specify for individual plumes values of temperature, depth of origin and composition.



CLOSER VIEW of the Loki lake is afforded by this high-resolution image, also made by *Voyager 1*. The large bright raft in the lake is traversed by several cracks, and pieces of bright material appear to have broken from its edges. They dot the surface of the lake. Perhaps the cooling crust of the lake has been disrupted by convective overturn or the addition of fresh eruptive material. Conceivably the raft is made of elemental sulfur. The lake is 250 kilometers across.

Additional information about the plumes can be derived from their distribution on the surface of Io and from their temporal behavior. At least nine plumes were seen to be erupting when *Voyager 1* arrived, and some of them were seen several times over the days of the encounter. The nine were more or less uniformly distributed in longitude but were concentrated at low latitudes: eight of them were within 30 degrees of the equator. Eight of the nine were active four months later, when *Voyager 2* arrived. (Pele, the largest plume, was no longer active.) Two of the nine formed a pair of plumes, one at each end of the fissurelike feature north of a black patch named Loki. They exhibited major variability in the height of their eruption during the encounters and also between them. Moreover, the surface patterns around the Pele and Loki vents changed substantially between the two encounters.

Pele and Loki are fairly near each other. In particular, they are both well inside a hemisphere intermediate between the one that faces Jupiter and the one that trails in Io's orbit. (Io, like the earth's moon, keeps the same face turned toward its planet.) That hemisphere has long been known to be much redder and darker on the average than the opposite hemisphere. The only other major changes on Io's surface between the two encounters also took place there. Two dark reddish ring-shaped deposits, each on the order of 1,400 kilometers in diameter, or about the size of the Pele deposit, developed around two calderas, both at high latitude, one, named Surt, in the north, the other, named Aten Patera, in the south. Neither was found to be erupting at the time of a Voyager encounter, but the activity at Surt may have been responsible for a period of increased thermal emission at a wavelength of five micrometers (in the infrared) observed at the Mauna Kea Observatory by William M. Sinton of the University of Hawaii at Honolulu one night about a month before the *Voyager 2* encounter. The Aten Patera deposit was discovered during a recent analysis of Voyager images by Alfred S. McEwen and one of us (Soderblom) of the U.S. Geological Survey.

The color pattern and the scale of these two enormous deposits make them closely resemble the one surrounding Pele; therefore McEwen and Soderblom suggest they belong to a class of large, short-lived energetic eruptions of which Pele is the best example so far. Employing measurements made by the infrared spectrometer on *Voyager 1*, John C. Pearl and his colleagues at the Goddard Space Flight Center of the National Aeronautics and Space Administration estimate that the temperature of the Pele vent is about 650 degrees Kelvin; Sinton's observations, presumably of Surt, suggest a similar figure.

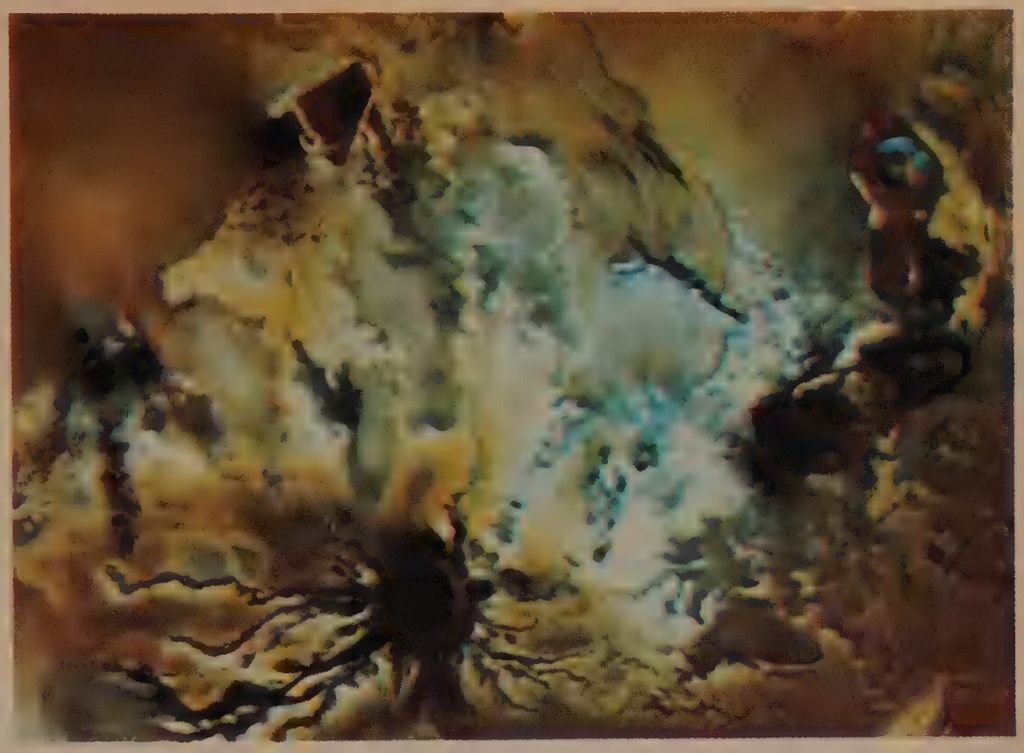
Eruptions of a second class are smaller: they deposit rings about 300 kilometers in diameter. They are long-lived: all of them were observed by both Voyager spacecraft, and from that fact the lifetime of such eruptions can be estimated as at least several years. They appear to be cooler: about 400 degrees K. They are restricted to an equatorial band of bright white material thought to be rich in sulfur dioxide frost. (Their ring-shaped deposits consist of the same material.) The best example is the feature called Prometheus, five degrees south of the equator. The pair of plumes at the Loki fissure seemed to oscillate between the two classes. The deposit surrounding them is complex: each has an inner zone consisting of a fan-shaped deposit, probably rich in sulfur dioxide, about 200



MAASAW PATERA and the surface markings surrounding that caldera on Io are much like the ones geologists see on the earth and on Mars. The principal difference is that Maasaw Patera is huge: the caldera alone is 50 kilometers across. Two levels in Maasaw Patera suggest stages of collapse. The greater part of the caldera is .7 kilometer deep, according to calculations based on the shadows in the caldera. The remaining part, at the upper right of the caldera, is two kilometers deep. At some point upwelling lava must have overflowed the rim of the caldera, producing streams of lava hundreds of kilometers long. The image is another of McEwen's *Voyager 1* mosaics.

kilometers long, and an outer zone similar in scale and color to the deposits around the three large, Pele-like plumes.

McEwen and Soderblom suggest that two distinctly different volcanic systems driven by different volatiles are responsible for the two classes. They also suggest that the classes arise from the peculiar relation in sulfur between temperature and viscosity. When sulfur is heated past its melting point, it undergoes several sequential changes. It starts as a yellow solid; then, at a temperature of about 400 degrees K., it melts to a low-viscosity yellow fluid. Further heating changes its color first to orange and then, at 430 degrees, to a clear pink. Next it turns deep red and thickens; by about 500 degrees it turns into a blackish tar. At about 600 degrees its viscosity starts to drop, and by 650 degrees it has become quite fluid again. Finally, at a temperature depending on the pressure, it vaporizes.



RA PATERA and the surrounding surface markings differ markedly from what geologists see on the earth and on Mars. Here Ra Patera is at the bottom of the image. It is a caldera filled with black material that sends out long, sinuous flows, mostly toward the left, over distances as great as 200 kilometers. David C. Pieri of the Jet Propulsion Laboratory of the California Institute of Technology and Carl Sagan of Cornell University propose that the dark center of Ra Patera represents a black, tarry phase of sulfur and that the flows represent cooler, less viscous phases. Two other calderas are notable. Toward the upper left an elongated caldera is banded in yellow, orange and black. The colors may represent different phases of sulfur. At the upper right a circular caldera includes a bluish crescent. In an image made by *Voyager 1* six hours before it made this one, the crescent is missing. Perhaps the crescent signals a low-energy eruption.

Low-viscosity molten sulfur that can readily transport heat occurs, then, in two forms: a clear reddish one at temperatures between about 400 and 430 degrees K. and an opaque black one at temperatures in excess of 650 degrees. The idea, therefore, is that the small plumes such as Prometheus are driven by liquid red sulfur coming in contact with liquid sulfur dioxide, making the sulfur dioxide vaporize and expand, as Smith, Shoemaker, Kieffer and Cook propose, whereas the large plumes such as Pele arise when hot silicates in the crust of Io vaporize sulfur at temperatures between 700 and 1,200 degrees K.

The remarkable difference in the duration of the two classes of plumes can be explained by this model. The small sulfur dioxide plumes persist for years because sulfur dioxide can be a liquid over a great range of depths in Io's crust. The liquid has extremely low viscosity (about that of alcohol at room temperature), and so it migrates easily through the crust to reach vents at the surface. In contrast, the large black sulfur plumes last for only days because the conditions favoring their eruption are delicately balanced. The venting of sulfur could stop, for example, if the temperature fell much below 650 degrees at the vent, in the conduit or deep in the crust where the sulfur and the silicates meet, because the sulfur would quickly cool into a tarry, nearly solid substance.

It remains an open question why the two classes of plumes are distributed as they are. The Voyager images of the dark red hemisphere of Io, where the large plumes are found, show numerous mountains that must consist of silicates protruding above Io's plains. (A sulfur mountain would not be strong enough to sustain its own weight.) Images of the opposite hemisphere have a lower resolution. In some of them, however, the region where the small, Prometheus-type plumes are most abundant is on the terminator (the day-night boundary), where mountains should be easily visible. In other images it is on the bright limb of Io. These views show mostly smooth plains.

Accordingly McEwen and Soderblom suggest there are major variations in Io's crust. Specifically, the crust in the region of the large, Pele-like plumes may be much thinner than elsewhere, so that hot silicates are closer to the surface and the pressure confining sulfur vapor is less. In addition the crust in the region of the small, Prometheus plumes may be enriched in liquid sulfur dioxide because some geophysical process (say a hydrostatic pressure gradient that varies with latitude arising from the rotation of Io) may drive the liquid toward the equator. It is worth noting that the Loki pair of plumes, which are perhaps a hybrid between the two distinct classes of plumes, lie in the area of overlap of the regions where the two classes are found.

A type of volcanic activity quite different from Io's geyserlike plumes is revealed by large volcanic calderas and their associated flows and surface markings, which are among the most prominent features identified in Voyager images. In general the calderas resemble terrestrial calderas except that they are larger: more than 200 of them are at least 20 kilometers across, as opposed to 15 on the earth, which has 3.5 times as much land area. The calderas on Io, unlike the plumes, show no strong concentration with latitude, being nearly as abundant per unit area at high latitudes as they are at the equator.

Debate about the nature of the calderas has centered on the issues raised by the plumes, namely the subsurface thermal conditions and the availability of sulfur and sulfur dioxide. Two extreme views of the calderas can be called the sulfur model and the silicate model. In the former, proposed in its purest form by Carl Sagan of Cornell University, the underground pools of molten sulfur postulated by Smith and his colleagues to explain Io's sulfur dioxide plumes also account for the volcanic flows on the surface; the flows around the calderas result from liquid-sulfur "lavas" rather than molten basalt or other silicates. In the latter, first proposed by Harold Masursky and Michael H. Carr of the Geological Survey and their colleagues, the calderas represent silicate volcanism much like that on the earth. The only difference is that the silicate flows on Io are tinted by sulfur.

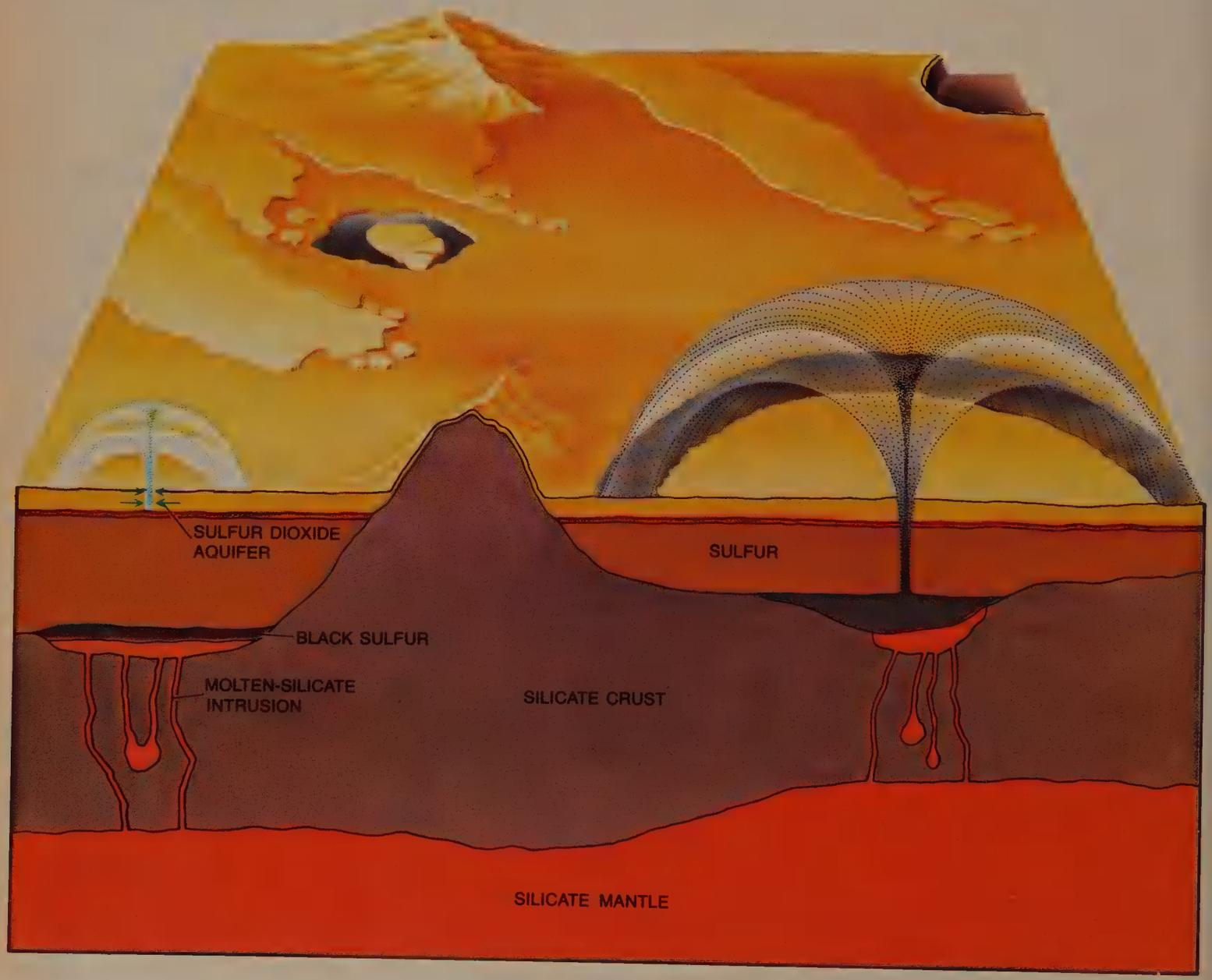
The sulfur model is supported by the undoubted abundance of sulfur and sulfur compounds on Io, by the detection of sulfur dioxide gas near one of Io's

eruptions and by the probability that these substances reach their melting point at shallow depths in Io's crust. In addition it is suggested that sulfur's sequence of changes in color and viscosity on heating or cooling explains the change in color with distance observed around many of Io's calderas. Ra Patera is a good example. David C. Pieri of the Jet Propulsion Laboratory and Sagan propose that the dark regions at the center of Ra Patera represent the high-temperature black phase of molten sulfur. Long, fingerlike flows that radiate from the caldera represent cooler, less viscous streams of the red and orange phases.

One problem with this proposal is that the behavior of the various sulfur phases under Io's conditions is uncertain. To be sure, the colors of the high-temperature phases can be retained in the laboratory by rapidly cooling the liquid. Nevertheless, the phases are only metastable and generally revert to ordinary solid sulfur after some period of time. Moreover, the phases on Io are likely to have impurities, and the impurities are likely to yield further complications.

The case for silicate volcanism rests in part on the measured topography of Io. The measurements are difficult because the spatial variations in the brightness of Voyager images resulting from differences in surface composition tend to be much stronger than the variations in brightness resulting from topographic relief. Still, some measurements have been tried on high-resolution images made under favorable illumination during *Voyager 1*'s closest approach to Io. The measurements suggest that Io has a range of topography far greater than that of the other Galilean moons: its mountains are from five to 10 kilometers high and at least a few of its calderas are from two to three kilometers deep. The maintenance of such a topography would be distinctly difficult on a body whose crust included thick deposits of warm, possibly liquid sulfur and sulfur dioxide. A stronger crust whose skeleton, at least, is silicate seems to be required; hence the broad distribution of calderas on the surface of Io would seem to rule out extensive subsurface sulfur "oceans."

The other main argument advanced by the proponents of silicate volcanism is that the structures seen on Io (calderas, long fingerlike flows, volcanic mounds with central craters and so on) are more or less what geologists are accustomed to seeing on the earth and on Mars. It seems implausible that sulfur, an unusual volcanic material with an exotic temperature-viscosity relation, should so closely mimic the structures produced by basaltic lavas. Unfor-



HYPOTHETICAL CROSS SECTION of Io shows an attempt to account for the body's volcanic activity. The mantle in the section is molten (or partially molten) silicate (*orange*); the crust is solid silicate (*brown*), overlain by layers of silicate debris rich in sulfur. The uppermost sulfur is cool and solid (*yellow*); then comes a thin zone in which the sulfur is molten (*red*) and a thicker zone in which the sulfur is hot and tarry (*yellowish brown*). At the left a plume like Prometheus is modeled. Molten sulfur makes contact with liquid sulfur dioxide (*blue*), which then boils up a conduit, producing at the surface a plume rich in sulfur dioxide snow. At the right a plume like Pele is modeled. Here a silicate intrusion heats sulfur so that it enters a very hot, tarry phase (*black*). Indeed, it vaporizes; hence vaporized sulfur erupts. A lava lake like Loki is at the upper left. A caldera is at the upper right. Its floor is the silicate crust underlying a thin layer of sulfur.

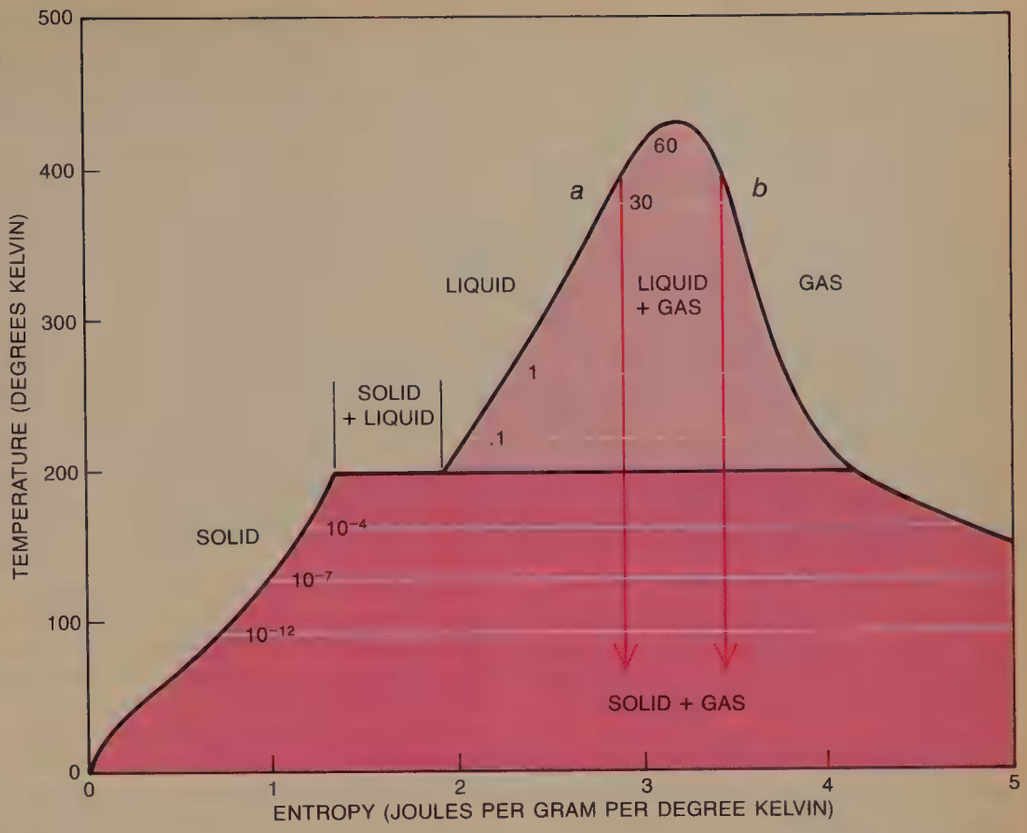
tunately geologists have only limited personal experience with sulfur as a volcanic fluid. There are rare sulfur flows on the earth, but they occur only when silicate volcanism melts sulfur deposits.

What, then, is happening on Io? Each side in the sulfur v. silicate debate now concedes the likelihood that the opposition's type of volcanic activity is occurring in at least some places on Io. Gerald G. Schaber of the Geological Survey observes, for example, that certain calderas are stamped out in mountainous regions of Io. Since the mountains presumably are silicate, the calderas must be silicate structures as well. Moreover, both sides agree that the volcanic activity at Io's surface, whatever its nature, is driven by molten silicates deep in Io's crust. The question is whether one type of volcanic flow, sulfur or silicate, dominates at the surface.

One further volcanic feature remains to be described. A discrete dark feature south of the Loki plumes is likely to be a huge lava lake of the kind that typically fills active calderas during eruptions on the earth. At the time of the *Voyager 1* encounter this feature was the largest "hot spot" on Io: it had a temperature of about 300 degrees K., whereas the local background temperature was only 130 degrees. Moreover, high-resolution images showed that the feature had a "raft" of bright material in it. The raft appeared to be traversed by cracks, and smaller pieces of the same material appear to have broken off from its edges. It is as if the cooling crust of the feature had been broken by convective overturn or the addition of fresh eruptive material. The feature is vastly larger than the lava lakes in Hawaiian calderas. Indeed, its length of about 250 kilometers would span the entire Hawaiian island chain. Is it filled with molten silicate or molten sulfur that now has chilled? We do not know.

The most striking evidence that the surface of Io is geologically young is the total absence of craters of impact origin in any Voyager image. Planetary geologists had found these signs of ancient and modern bombardment everywhere else they were able to look, including Mars and the earth, and the cratered surfaces of Ganymede and Callisto show that the Jovian satellites have had a history of impacts similar to that of the inner planets. Furthermore, calculations by Shoemaker suggest that the current flux of comets and asteroids through the Jovian system should be producing large craters on Io at a rate within an order of magnitude of the rate at which the earth's moon is now being cratered.

On the basis of Shoemaker's finding we have calculated how much volcanic activity is needed to bury Io's craters. We conclude that a one-millimeter thickness of volcanic debris must be spread over Io in an average year. This



THERMODYNAMICS OF SULFUR DIOXIDE suggest that a range of geyserlike eruptions is possible on Io. The vertical scale plots the temperature of sulfur dioxide; the horizontal scale plots the entropy. (Pressures are indicated by isobars labeled in units of bars.) On such a plot the horizontal position of a parcel of sulfur dioxide represents proportions of phases (for example liquid and gas). Moreover, a vertical line represents the evolution of a sulfur dioxide eruption that does no work (such as pushing against a barrier) and neither gains nor loses heat. Two such eruptions are shown. One of them (*a*) begins when liquid sulfur dioxide is raised to 393 degrees K., the temperature of molten sulfur deep in the crust of Io. The sulfur dioxide thus boils upward toward the surface. The other eruption (*b*) begins with gaseous sulfur dioxide. In each case the sulfur dioxide explodes as a cloud of ice and gas in the near-vacuum above Io's surface. The thermodynamics were analyzed by Susan W. Kieffer of the U.S. Geological Survey.

implies a very active body indeed. Specifically, it implies that Io erupts several thousand metric tons of material per second, or about the total amount of material ejected by the Mount St. Helens eruption of May, 1980, every month. The rate has significant consequences. Among other things, it suggests that at least the upper part of Io's mantle and crust has been recycled many times in the span of Io's history. This recycling is consistent with the idea that Io has been thoroughly "sweated out" by tidal heating, so that it has lost whatever light volatiles it once had.

The details of the recycling are not well understood. For example, the relative importance of plumes and lava flows for covering the surface with new material is not clear. Estimates of the total amount of material erupting in Io's plumes imply that the resurfacing rate attributable to plumes alone ranges from a few ten-thousandths of a millimeter per year to a tenth of a millimeter or more. This suggests that flows are at least competitive with the plumes. Of course, one must bear in mind about such estimates that Io is highly dynamic and that the Voyager encounters with it have given us, so to speak, only two snapshots.

Is any evidence of Io's volcanic activity apparent from the earth? Remarkably, the answer is yes. In fact, evidence of volcanoes on Io existed in puzzling data gathered years before the Voyager encounters. Telescopic observations of Io at infrared wavelengths, made in the late 1960's and early 1970's, reveal two curious things about the object. First, when Io is eclipsed by Jupiter its temperature falls precipitously, as one would expect of a body with little or no atmosphere whose surface is a layer of fine insulating powder. On Io, however, as opposed to the earth's moon, the minimum temperature, attained when Io is deep in Jupiter's shadow, is too high to be consistent with a homogeneous insulating layer. Second, the temperature of Io inferred from its infrared brightness is not the same for all infrared wavelengths; it is significantly greater at shorter wavelengths.

Individually these problems could be handled, although in an ad hoc fashion. Two-layer models of Io were devised to explain the eclipse data, and wavelength-dependent variations in Io's emissivity were proposed to explain the

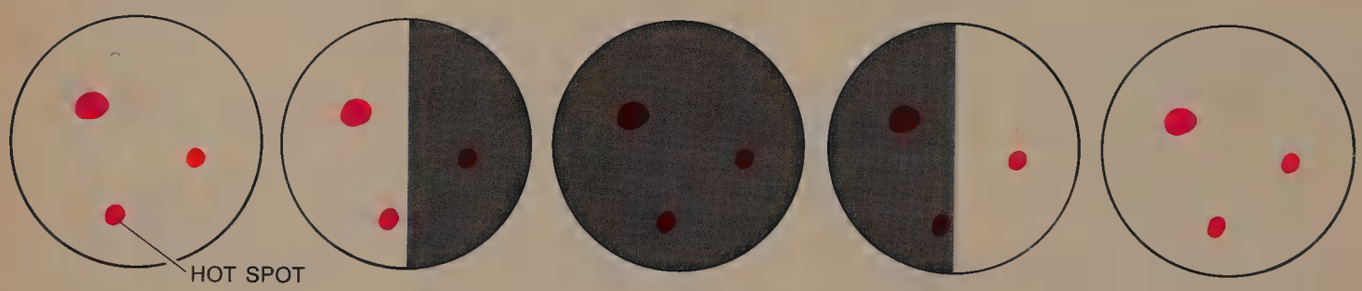
temperature anomalies. After the Voyager encounters and the discovery of volcanoes on Io, three investigators at the Jet Propulsion Laboratory, Dennis L. Matson, Gary A. Ransford and one of us (Johnson), returned to the puzzling data, hoping to solve the old problems in the light of the new information. They devised a simple model in which the insulating surface of Io is dotted with hot spots of a given temperature. It emerged that the problems would disappear if about 1 percent of Io's surface were covered with hot spots whose temperatures are similar to the ones the Voyager spacecraft measured.

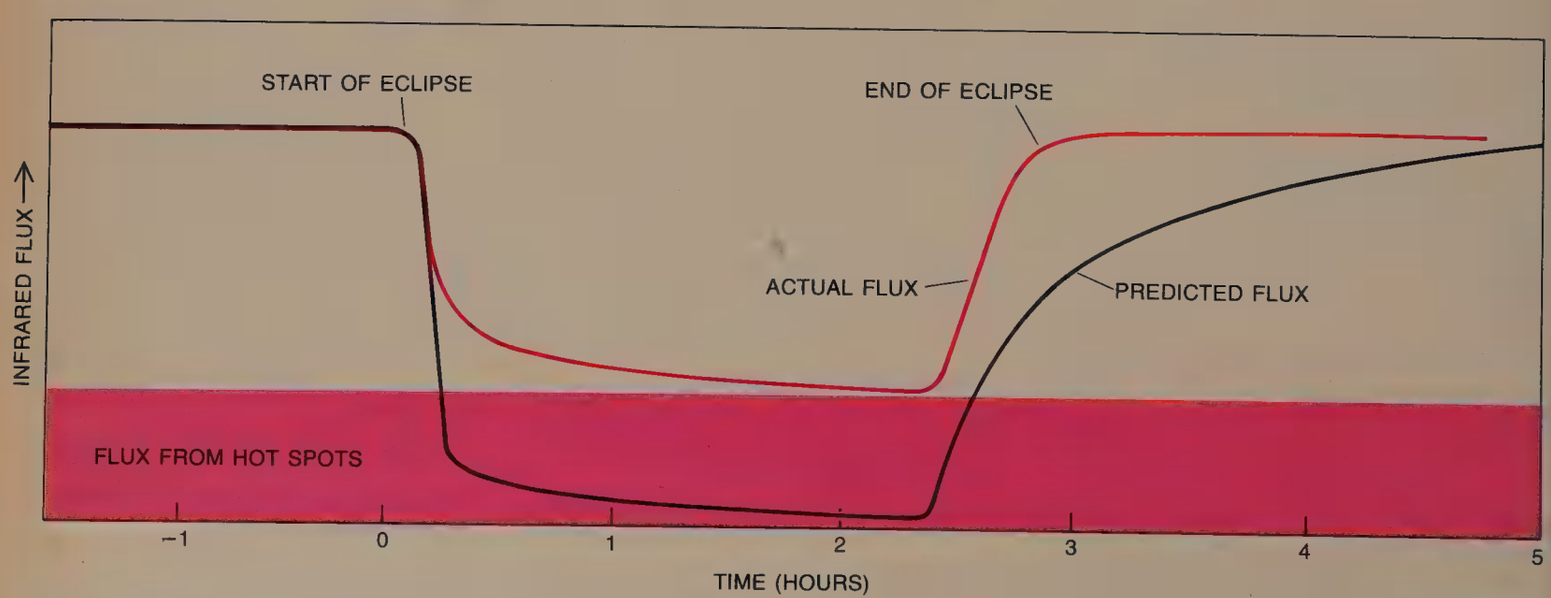
The modeling yielded a value for the amount of energy radiated by the hot spots: averaged over the surface of Io, it came to approximately two watts per square meter. New analysis of Voyager data and new infrared observations made from Mauna Kea yield independent estimates of the amount; they agree with the model and with the earlier data. The current best estimate of the average heat flow from Io, drawing on all these sources, is 1.5 watts per square meter, plus or minus .5 watt. In comparison, the earth radiates about .08 watt per square meter, and the moon, with nearly the same size and mass as Io, radiates only .03. The radiation from Io stands out as being completely anomalous. Even without the proposal of tidal heating it would be clear that a source of energy other than internal radioactivity is active on Io.

It should be said that the heat flow on Io differs from the heat flow on the earth and the moon not only quantitatively but also qualitatively. On the earth and the moon nearly all the excess internal heat generated by radioactivity reaches the surface by conduction through the crust. (Volcanic eruptions on the earth account for only a small part of the total.) From there it is radiated into space without measurably raising the surface temperature of the body. (The surface is heated primarily by the energy the surface gets from the sun.) This means the excess cannot be measured at a distance as infrared emission. What is required instead is the precise measurement of the thermal gradient in the upper part of the crust. On Io matters are different. The excess internal heat reaches the surface mostly by convection: it is carried by hot fluids rising to the hot spots. From there it is radiated into space at temperatures far higher than Io's surface temperature; hence it is readily measured. The energy output from Io's surface is between 10^{13} and 10^{14} watts.

So far two processes other than internal radioactivity have been proposed as the sources of this energy. One is tidal heating; the other is that Io is heated by its resistance to electric currents driven into the body by its interaction with Jupiter's magnetosphere. Indeed, Thomas Gold of Cornell suggests that Io's plumes arise from such an interaction when ionized matter ascends along the field lines of current flowing into Io's surface like great sustained lightning bolts. He bases his suggestion on the difficulty of driving volcanic gases to velocities exceeding the speed of sound and on data from the Voyager magnetometers implying that a current of about one million amperes flows along magnetic field lines in Io's vicinity. (A current of one million amperes is also what theoretical calculations of the flow of ions in the Jovian magnetosphere would lead one to expect.)

Io's output of 10^{14} watts can be exploited to test the possible importance of electromagnetic effects in the object's overall energy balance. Suppose all the current thought to impinge on Io is converted into heat in the interior by resistive heating. The net amount of heating would range from 10^{11} to 10^{12} watts, no more than a hundredth of the amount needed to account for Io's output. Electromagnetic effects might nonetheless be important for events such as eruptive plumes. In that case, however, the "lightning bolt" giving rise to a plume would produce a spot on Io with an effective temperature on the order of 100,000 degrees K. Searches for such spots on the night side of Io have not been successful.





ECLIPSE COOLING of Io puzzled observers who measured the infrared flux from Io as Io passed through Jupiter's shadow. The observers thought the flux would fall rapidly toward zero at the start of each eclipse and then return to its normal value when sunlight warmed Io's surface (*black curve*). Instead the flux stayed well above zero (*colored curve*). The explanation is that volcanic "hot spots" on Io emit a more or less constant flux of heat throughout the eclipse. Io's volcanic activity can therefore be monitored from the earth.



TORUS OF IONIZED SULFUR circles Jupiter and is centered on Io's orbit. It derives from Io's volcanic activity, which is thought to eject about one metric ton of sulfur and oxygen atoms per second into the space surrounding Jupiter. This image of one limb of the torus was made at the Mount Wilson and Las Campanas Observatories by John T. Trauger of Cal Tech. False colors designate the proportions of singly ionized sulfur atoms (*blue*) and doubly ionized sulfur atoms (*red*). The apparent right edge marks the limit of Trauger's data.

Meanwhile a problem has arisen with the tidal-heating hypothesis. The tidal interaction of Io and Jupiter involves tidal bulges raised on both the satellite and the planet; the bulge on Jupiter applies a gravitational torque to Io, causing the satellite to accelerate and move into an ever higher orbit. (A similar interaction is responsible for the slow but steady retreat of the moon from the earth.) The current best estimates of the strength of the interaction place an upper bound on the average input of energy to Io, and that upper bound is about a factor of two lower than the measured heat flow from Io.

The problem is still unresolved. Perhaps Io's heat flow and volcanic activity are highly variable over time, and the Voyager spacecraft arrived in a period when the activity was unusually pronounced. Perhaps some of the parameters and assumptions involved in the calculations of Io's orbital evolution and tidal heating are mistaken. Or perhaps the hot spots are not uniformly distributed over Io's surface. Recent data gathered at NASA's Infrared Telescope Facility in Hawaii by workers at the Jet Propulsion Laboratory and the University of Hawaii including one of us (Johnson) suggest that the region around Loki may contribute a large fraction to Io's total heat flow. If that is the case, and if Loki and its vicinity was mistakenly taken to be representative of the entire body, the total power actually emitted by Io may be closer to the theoretical expectation.

There seems to be no evidence that Io's heat flow varies much: the eclipse data recorded in the early 1970's give results in agreement with Voyager measurements and with eclipse data recorded in the 1980's. One type of infrared emission does exhibit dramatic short-term changes: the emission near a wavelength of five micrometers shows occasional bursts that probably are related to day-to-day processes on Io. The first such burst was detected in 1978 by Fred C. Witteborn of the Ames Research Center and his colleagues aboard the instrumented high-altitude aircraft named the Kuiper Airborne Observatory. Their brief observation of a five-micrometer flux of more than twice the intensity of what could be due to sunlight reflected from Io was startling.

The flux suggested that a small area amounting to 10^{-4} of Io's surface had a temperature of about 600 degrees K., but volcanic activity seemed almost unthinkable at the time. In the wake of the *Voyager 1* encounter the significance of Witteborn's observation was immediately apparent: the infrared spectrometer on *Voyager 1* had detected small regions on Io that had a similar temperature. William Sinton, working at Mauna Kea, began his systematic search for more "five-micron events." One of his first successes was the detection of the event between the Voyager encounters: the event thought to have been associated with the change around the Surt caldera.

Since then Sinton has accumulated several years' worth of data. He has found few events that rival the early ones, but his data show a continual variation in the flux from Io's very hot "hot spots." The total amount of energy these small, high-temperature areas radiate is much less than the amount more extended areas emit at lower temperatures. Hence the five-micron events do not contribute significantly to Io's overall energy budget. They nonetheless yield insight into the violent events that presumably generate large, Pele-like plumes. In addition they will facilitate the search for evidence of silicate lavas. More broadly, observations of Io made from the earth, including eclipse studies and infrared measurements, are all potential ways of monitoring the volcanic activity on Io and will provide an important link between the Voyager data and the data to be amassed by the next spacecraft sent to the Jupiter system.

The next spacecraft will be *Galileo*. Scheduled for launch in 1986 and arrival near Jupiter in August, 1988, it will include a probe to enter the atmosphere of Jupiter and a long-lived orbiter to allow some 20 months of detailed observations of the planet, its magnetosphere and its satellites. On its way toward Jupiter *Galileo* will pass within 1,000 kilometers of Io, or 20 times closer than the nearest approach of the Voyager spacecraft. Instruments aboard *Galileo* will map Io's surface and make images at a resolution comparable to that of Landsat images of the earth. The interaction of Jupiter's magnetosphere and Io's tenuous atmosphere will be examined, and radio tracking of the spacecraft will place new constraints on models of Io's interior.

After this nearest encounter *Galileo* will have to avoid the vicinity of Io: it would otherwise be disabled by a prolonged dose of radiation. Still, most of the orbiter's 20-month mission will keep it well within the range from which the Voyager spacecraft detected Io's volcanoes. And on about a dozen occasions *Galileo* will be able to scan Io for volcanic changes, make thermal maps that show the location of hot spots and search for evidence of ions passing from Io into Jupiter's magnetosphere. Io had already intrigued astronomers for 400 years when the Voyager spacecraft revealed a world more complex and yet more fascinating than it had been before. The trend will surely be continued by the next round of exploration.

An Early Iron Age Farm Community in Central Europe

Excavations at a site in Bavaria unearth clues to an economy of 1000 to 800 B.C. in which farmers were just beginning to exchange their surplus for the work of specialized craftsmen

by Peter S. Wells

Town life was commonplace in the Near East by 3500 B.C., but most settlements in Europe north of the Alps were little more than hamlets until the end of the late Bronze Age and the beginning of the early Iron Age. In that part of the world the transition to town life first took place in about 800 B.C. It was a time of rapid change in temperate-climate Europe that saw both the growth of trade and the expansion of metal production. Archaeological knowledge of the transition has come mainly from the excavation of cemeteries that number in the thousands and from the chance (but not infrequent) discovery of buried hoards of metal. Only a few European settlements of the period, however, have been investigated systematically, and so the economic developments that led to the formation of the first central European towns remain rather poorly known. Here I shall describe the findings of four seasons' work at a farmstead site of the period in Lower Bavaria and relate those findings to the rise, both there and elsewhere in central Europe, of larger settlements: precursors of the trading towns of the medieval period.

The site, Hascherkeller, is on a sand and gravel terrace that forms the northern border of a narrow river valley on the outskirts of Landshut, the principal city of Lower Bavaria. The terrace is 15 meters above the river (the Isar, a tributary of the Danube), and its waterborne deposit of glacial debris is covered by a thick layer of loess, a pale yellow sedimentary soil attributed to wind deposition at the end of the Ice Age. The loess in turn is covered by a rich humus, the product of millenniums of soil development, and the top 40 centimeters of the humus has been disturbed by modern deep plowing. Hence nothing of the original ground surface of the prehistoric settlement survives. Only those features that were a result of the settlers' own deep digging, such as pits of various kinds and boundary ditches, still remain: dark, humus-filled intrusions in the otherwise undisturbed loess.

Beginning in 1978 my colleagues and I set about dividing the area of intended excavation into mostly adjacent five-by-10-meter plots. The plow-disturbed humus was removed with shovels in two consecutive excavations, each to a depth of 20 centimeters. The archaeological material in the disturbed humus was collected and catalogued according to the level and the plot of origin. Once the bottom of the disturbed humus was reached the work proceeded with hoes and trowels until the top of the underlying loess came to view. Before the intrusive features thereby uncovered—the pits and ditches—were further excavated they were mapped and photographed. The humus fill in each pit was then bisected and the two halves of the fill were examined separately. All the fill we removed was screened through quarter-inch wire mesh.

As we proceeded it became apparent that the settlement had consisted of three enclosures side by side and extending from east to west, each one bounded by a double ditch. The enclosures were well defined on their northern, eastern and western sides, but stream erosion had destroyed their southern side. Our work in the first year and in the seasons that followed concentrated on the contents of the western enclosure and the middle one.

Although plowing had destroyed the top of the double ditches, what remained was as much as three meters wide. Excavation showed that the ditches were *V*-shaped and about 1.5 meters deep. The fill was dark brown and contained small fragments of pottery and animal bone. One section of the inner ditch of the western enclosure had a row of 19 postholes at the bottom; evidently a wood palisade had once stood there. The average distance between the centers of the postholes was 13.7 centimeters. If this figure represents the average diameter of the individual posts, the palisade was a stout one.

Experiments have indicated that earth-set posts of this kind rot within a few decades. Even though no other postholes have been found so far, many sections of the perimeter ditches show evidence of redigging, which would have been necessary to set up new posts as the old ones decayed. This suggests that all the ditches held palisades. Their purpose was presumably not so much to guard against possible attackers as to keep domestic animals inside the settlement and to keep wild animals out at night.

Most of the information about the community at Hascherkeller has come from the contents of the pits found within the enclosures. Of the 21 largest pits we excavated, 11 could be assigned on the basis of their contents to the three-enclosure settlement; the remaining 10 belonged either to earlier and smaller Bronze Age settlements or to later Roman occupations. Like the ditches, the pits with their dark fill stood out sharply against the loess subsoil. Indeed, the fill of the pits was even darker than that of the ditches, an indication that the pits had held additional organic matter.

The form, size and contents of the 11 pits put them in five functional categories. Three of the pits were long, narrow and cup-shaped, and they were oriented toward the cardinal points of the compass; two were oriented north-south and one was oriented east-west. Such orientations have been found to be prevalent among houses of the same period



RECTANGULAR EXCAVATIONS exposed one of three farmsteads at Hascherkeller, an Iron Age site in Lower Bavaria. The discoloration of the soil just in front of the two-meter scale rod reveals part of the ditch system that separated this farmstead from its neighbor to the west. The farmers erected palisades in the ditches in order to keep livestock from straying and to keep wild animals out at night.



PROFILE OF A DITCH dug at Hascherkeller three millenniums ago is marked by a *V*-shaped intrusion of dark humus into the yellowish loess subsoil covering the gravel of a river terrace. The ditch was one of two concentric ditches surrounding the westernmost of the farmsteads. In the ditches were found fragments of bone, pieces of baked mud plaster from the building walls and sherds of pottery.



SITE OF THE FARMSTEADS is above the Isar River on the outskirts of the city of Landshut. Nearby are Hallstatt, an Iron Age industrial site, and The Heuneburg, a trading town.

excavated elsewhere in Europe. The coincidence suggests that the pits served the function of "cellar holes" for houses. They were probably storage places for large pottery vessels that protected foodstuffs from dampness, temperature fluctuations and animal predation. The potsherds found in the three pits support this conclusion; most of them were fragments of thick-walled vessels.

The mouth of three of the deepest pits was circular and the walls were nearly vertical. Similar pits at other settlement sites of the period have been found to contain carbonized grain; moreover, the pits had been lined with basketry or clay, indicating that they had served for grain storage. The three pits at Hascherkeller did not yield any such positive clues to their use. They contained few potsherds or other kinds of waste, however, and may well have served a similar purpose.

Two more pits seem to have been associated with metallurgical activity. One of them held a hammerstone and a sandstone mold for casting finger rings. Several small scraps of bronze were found nearby. An adjacent pit contained red-stained pebbles and soil, apparently discolored by intense heat. One may conjecture that a hot fire in the second pit melted down bronze for casting. The pit where the mold was found also yielded a loom weight made out of fired clay and five clay spindle whorls. These suggest that the area served for weaving as well as casting.

One large pit, some meters north of the central enclosure, contained much charcoal, and at its bottom were the remains of a boxlike clay structure. More than half of all the potsherds found at the site were taken from this pit. Most of them were from what potters call wasters: pottery that had burst or warped as it was being fired and was therefore discarded. The conclusion seems inescapable that the clay structure in the pit was the firebox of a potter's kiln.

The last two pits, shallow and with gently sloping sides, held small fragments of pottery and animal bone. Apparently these shallow excavations had been borrow pits: places where the Hascherkeller inhabitants dug loess as a raw material for making their pottery or for plastering the walls of their houses with mud. Thereafter the hollows had gradually filled up with settlement debris. In this connection, the 11 pits yielded a total of 198 kilograms of baked mud, fired either when a structure had burned down or when the mud had been plastered on a wall adjacent to some source of heat, such as a hearth. The rest of the mud had simply weathered into silt again after the settlement's buildings had fallen into disuse.

The most abundant evidence of human occupation at Hascherkeller was the broken pottery. The total number of sherds recovered was 14,853. The great majority of them were found in the pits, but 3,828 came from the ditches, from the humus layers overlying the loess subsoil and from other areas. All represent plain, coarse wares typical of ordinary farmers' pottery. Many graves of this period in central Europe contain fine decorated wares, but fewer than 4 percent of the sherds from Hascherkeller had decorations of any kind.

We sorted the sherds on the basis of their thickness and found that they fell naturally into three categories. The first category consisted of relatively thin-walled wares; the sherds were less than 4.5 millimeters thick. The second category consisted of sherds from 4.5 to nine millimeters thick, the third of sherds more than nine millimeters thick. The remnants of small cups, beakers and bowls fell into the first category; they were also the fewest in number and the most frequently decorated. Sherds from larger bowls and from high, wide-mouthed jars with a coarse surface finish made up the second category. Those representative of the third category were chiefly from rough-surfaced jars that probably served mainly for food storage. At other habitation sites of the period many such intact vessels, buried in storage pits and cellar holes, have been found with grain still in them.

Except for the great quantities of fired mud, which provided useful information about the location of structures in the settlement, the most numerous remains at Hascherkeller were fragments of animal bone, 1,435 of them. Brenda Benefit, a doctoral candidate at New York University, has analyzed the bone fragments and finds that 253 of them can be identified as belonging to specific parts of animals of known species. The identifiable bones are predominantly (87 percent) those of domestic animals; the rest are the bones of wild animals. Of the bones of domestic animals, those of pigs predominate (37 percent), those of cattle come next (24 percent) and those of sheep and goats together account for almost all the rest (33 percent). Also found were fragments of a small number of horse and dog bones. The wild animal principally represented among the bone fragments is the red deer (*Cervus elaphus*), but there are also fragments of the bones of hares, hedgehogs and one species of bird (the quail). There are in addition a good number of fishbones but none that allow the identification of the species.

Benefit has analyzed the animal teeth uncovered at the site. She finds that the wear of the pigs' teeth has a bimodal pattern, indicating that the animals were slaughtered at two different times in their life span. Many had been eaten as suckling pigs, not long after birth. The rest were killed at about age two. This is a pattern of pig slaughtering common in Europe down to the present day. It maximizes the yield in meat with respect to the quantity of feed that must be given

the animals when in the winter months they cannot forage for themselves.

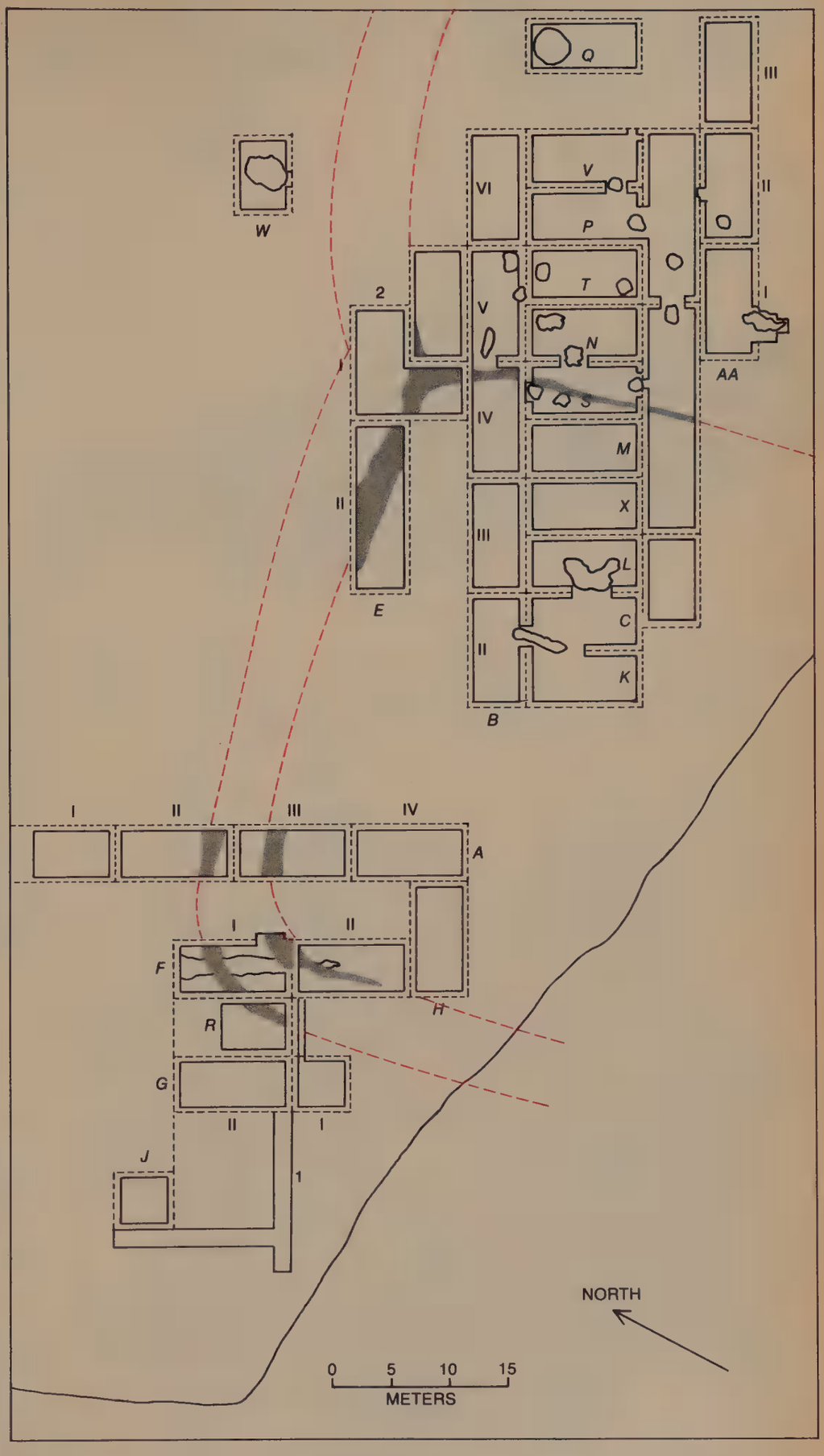
Sheep, goats and cattle led longer lives, regardless of the cost in winter fodder. Presumably the sheep were raised mainly for their wool and the goats and cattle for their dairy products. The cattle may also have been valued as draft animals and, when they were eventually slaughtered, as a source of hide.

In addition to this substantial array of meat and dairy resources the remains of various plants show that the farmers of Hascherkeller relied on cereals, garden crops and certain wild foods to augment their diet. Analysis of the plant remains by Caroline Quillian Stubbs, a doctoral candidate at Harvard University, indicates that the main cereal crops were millet, wheat and barley and that lentils were also cultivated. The inhabitants collected hazelnuts and the products of several other wild plants that are looked on as weeds today but that played an important role in the diet of earlier Europeans. At this settlement the inhabitants collected cleavers (*Galium*), goosefoot (*Chenopodium*) and sorrel (*Rumex*).

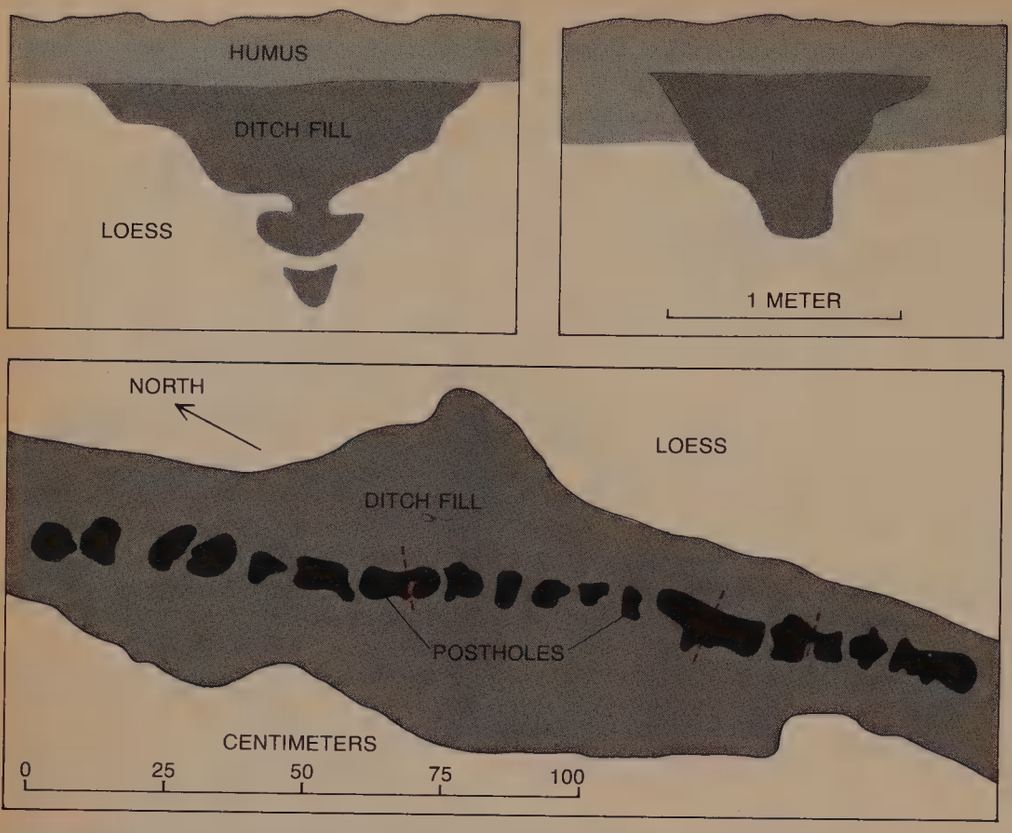
What picture of early Iron Age life can one build from these humble clues? The estimated size of the three ditched enclosures that make up the Hascherkeller farmsteads, each consisting of some 3,000 square meters, corresponds to the size of individual enclosed farmsteads at many other late prehistoric and early historic settlements in Europe. It is likely that each farmstead was inhabited by a family numbering from five to 10 individuals and that each included a dwelling, a barn for the stock and smaller structures such as sheds and workshops. It is evident that all three farmsteads were operating at the same time: their perimeter ditches meet neatly and never cut across each other.

The daily life of the 15 to 30 men, women and children in the settlement can be considered under three interrelated economic headings: subsistence, manufacturing and trade. Under the first heading there is abundant evidence for a self-supporting economic organization: animal husbandry that yielded both meat and dairy products supplemented by hunting and fishing, along with cereal and legume production supplemented by the collection of wild foodstuffs. Whether or not these farm activities yielded a surplus of meat or grain, it is probable that such activities as cheesemaking and hide dressing furnished the settlement with easily preserved commodities in excess of the farm families' own needs.

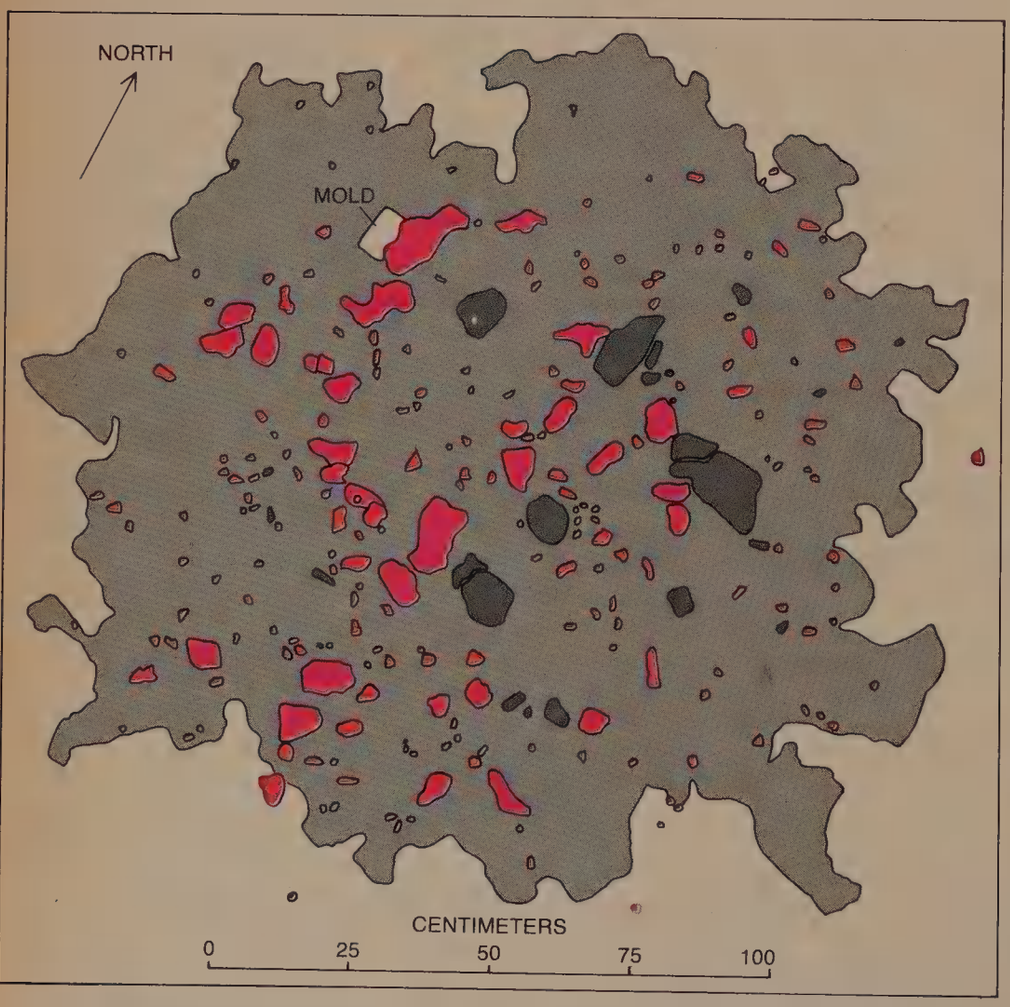
What commodities in addition to cheese, leather and possibly meat may have been produced in surplus quantities? They did not include pottery. At



TWO OF THE THREE FARMSTEADS, the westernmost and central ones, appear in this site plan. The excavated parts of the double ditches that surrounded each farmstead are in gray; their further extent, as indicated by a magnetometer survey, is outlined with broken colored lines. Of the 21 larger pits (*solid outlines*) uncovered by excavators, 11 had been dug during the early Iron Age. More than half of the pottery fragments found at the site came from a single pit (*W*) beyond the ditches of the central enclosure, which also contained the remains of a pottery kiln. Two cellar-hole pits (*C, AA I*) led the others in the quantity of mud-plaster fragments they contained, respectively 75 and 44 kilograms. The west pit in rectangle *N* (*see bottom illustration on next page*) held a quartzite hammerstone, a sandstone mold for casting bronze finger rings, a clay loom weight, five spindle whorls and 20 kilograms of plaster fragments.



DITCHES WITH A *V*-SHAPED CROSS SECTION (*left and right at top*) are distinguished from the lighter loess subsoil by the darkness of their humus fill. The plan view of a section of one ditch (*bottom*) shows 19 postholes over a distance of 1.6 meters. The broken lines in color indicate double holes. This section is a part of the inner ditch of the western enclosure.



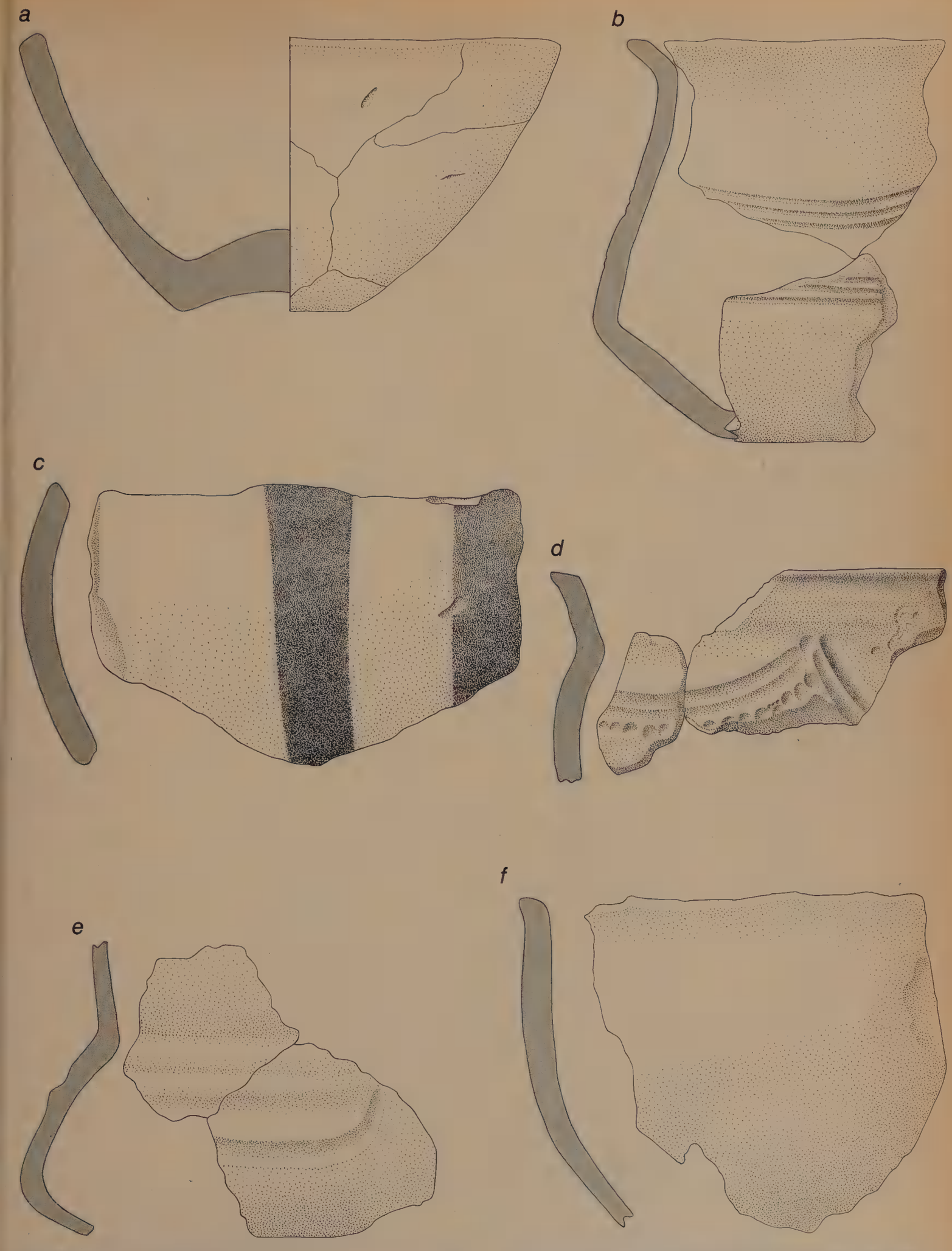
"WORKSHOP" PIT in rectangle *N* contained, in addition to casting and weaving equipment, numerous potsherds (*black*) and plaster fragments (*color*) in sufficient quantities to suggest that the structure housing the metal and textile workers also had mud-plastered walls. The sandstone casting mold is shown in situ to the left of center near the top of this plan view.

Hascherkeller each clay bowl, cup or jar was made not on the potter's wheel but by laborious coiling and paddling, which is unlikely to have given a farm settlement a pottery surplus. Our discovery of the loom weight and spindle whorls, together with the evidence that sheep were kept for a number of years before they were slaughtered, suggests, however, that one commodity perhaps made in excess of local needs was wool cloth.

This brings us to the third economic heading: trade. Evidence that the settlement's inhabitants imported exotic materials includes the fact that bronze artifacts such as finger rings were being made locally out of melted-down scrap bronze. The scrap was certainly imported, and several of the bronze pins unearthed may also have been brought in from the outside. On the other hand, we also found five fragmentary iron objects, too small and corroded to identify. Although they vouch for the settlement's being one of the Iron Age, they do not necessarily indicate that the farmers imported any iron artifacts. Iron ore is nearly ubiquitous, whereas the raw materials for the alloy we call bronze, tin in particular, are not. Evidence for local iron smelting, in the form of one lump of iron slag, supports the conclusion that the people of Hascherkeller were familiar with ironsmithing. There is no evidence, however, for local bronze smelting.

Additional imports include glass beads and graphite, the latter used for the surface decoration of a few of the local pots. If a smear of graphite is applied to the vessel's surface before firing, a shiny black finish is formed. This kind of ornamentation, either as a complete surface coating or as a series of bands, was particularly popular in central Europe beginning in about 1000 B.C. The principal sources of the graphite were deposits at least 100 kilometers from Hascherkeller, east of Passau on the Danube and to the north in Bohemia. As for the glass beads, four of them, blue green in color, were among the artifacts recovered: two from the pit that held the pottery kiln and two from another pit. Their place of manufacture is not known, but no evidence for local glass production has been found either at the settlement or at other sites of the period in central Europe. Judging from the relatively wide distribution of such beads, however, it seems likely that more than one bead-production center was regularly exporting its wares to the farm communities of the region.

What makes this pattern of scrap-bronze, graphite and bead importing particularly significant in relation to the later development of Iron Age towns in Europe north of the Alps is that all three



SHERDS FROM SIX POTS are fitted into the complete or partial profile of the original vessel. The decorated sherds are thinner than one of the two undecorated sherds (*a*). One vessel (*c*) was decorated with black bands, made by rubbing its surface with graphite before it was fired. The nearest sources of graphite were 100 kilometers away, so that the local potter probably obtained the material by trading.

imports were luxury products. The manufacture at Hascherkeller of metal sickles or other agricultural implements, although they can perhaps be regarded as overly stylish, has a utilitarian purpose. No farmer, however, really needs bronze rings and pins, graphite-ornamented pottery or blue green beads in order to raise more pigs or sow more millet. The settlement was trading with the larger world not for the necessities of life but for ornaments.

When was all this going on and for how long? Six charcoal samples taken from three pits yielded carbon-14 dates indicating the time of occupation was a 200-year period between 1000 and 800 B.C. The settlement's pottery and bronze objects and the imported glass beads in turn fit with a central European archaeological phase known as Hallstatt *B*. When material from this phase is cross-dated with material from Mediterranean cultures for which historical records exist, the Hallstatt *B* phase is assigned to the three centuries between 1000 and 700 B.C. Thus both the absolute chronology and the relative chronology are in close agreement.

As for how long the settlement endured, here the evidence is less direct. The ditches were renewed, and I have suggested that the renewal was needed in order to rebuild rotted palisades. Since the timbers of the palisades would have lasted for a few decades, the settlement could have lasted no less. Another indication of the length of occupation comes from the mud plaster preserved by baking. Many pieces of baked mud give evidence that the walls of some structures had been replastered and repainted two or three times. This suggests occupation for as long as two or three generations. Could the occupation have been even longer? Both the carbon-14 determinations and the analysis of the pottery indicate an upper limit of no more than two centuries of occupation. The pottery is completely homogeneous in style. It is difficult to believe the settlement would have remained unaffected by the winds of change over any period longer than 200 years.

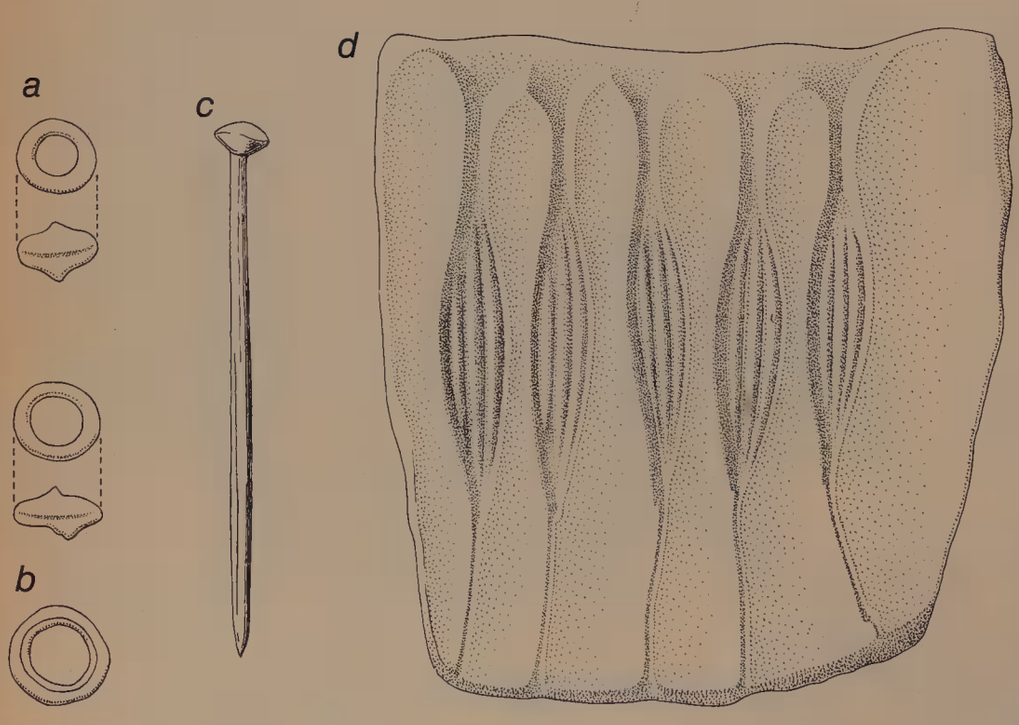
The picture of Hascherkeller that emerged after four seasons of excavation is one of a community economically self-sufficient for its necessities but at the same time tied into the larger world of central Europe by the desire for small luxuries. To satisfy these desires only one path could be followed: trade. Trade, however, is a two-way path. What the farmers could offer their trading partners was farm produce: wool cloth or yarn, cheese, possibly butter and salted meat and hides (or finished leather goods).

At this time a general intensification of agricultural production was under way in central Europe. Whether or not the Hascherkeller settlement remained active for a full two centuries, other farm settlements were then being occupied for as long and much longer. Manuring, fallowing and crop rotation were preserving soil fertility. Many more metal tools were being manufactured, not only sickles but also axes, saws, chisels and hammers, all implements that increased human efficiency.

Imagine for a moment that expanded metallurgy had led to the development of associations of smiths who did no farming of their own. As the graves of the period reveal, more than metal implements were being cast and forged. The dead were accompanied by ornate weapons and household goods: swords, helmets, large bronze vessels and gold ornaments. How would the smiths have fed themselves? Probably by trading their manufactured goods for surplus agricultural produce from the increasingly productive rural farmsteads. One can even imagine traveling middlemen finding their place in such trade networks.

The archaeological record provides an actual example of such an association that arose in what is now Austria no more than 160 kilometers from Hascherkeller. At the salt mines of Hallstatt a community with a population of some 200 between 800 and 400 B.C. devoted its energies exclusively to the extraction and trading of salt. The exceptionally rich assemblages of trade goods that accompanied the Hallstatt dead to the grave are eloquent evidence of the success of their experiment in a one-commodity communal venture. The first salt miners at Hallstatt began their work in about 1000 B.C., but they may have been farmers who mined salt chiefly for themselves, as the Hascherkeller bronze casters, weavers, shepherds and cheesemakers were to do in their own small settlement in the centuries that followed.

In any event the rise of the mining town of Hallstatt was not unique. Towns with populations in the hundreds, busily devoted to smelting and forging iron, sprang up in the Alpine border region of what is now Slovenia. Closer to Hascherkeller, in southwestern Germany, one well-studied site, The Heuneburg, grew into a commercial center: a densely packed cluster of substantial wood buildings not unlike the early medieval trading towns of 14 centuries later. The craftsmen of The Heuneburg engaged in different kinds of primary manufacturing rather than in some single specialty, and the same was true in similar commercial towns in central Europe. The principal function of the towns, however, was to engage in trade. None of these population clusters could have arisen without the support of thousands of small agricultural settlements such as Hascherkeller, able and willing to produce ever greater agricultural surpluses to exchange for the townsmen's goods. In Shakespeare's *The Tempest* Antonio says, "What's past is prologue." At Hascherkeller we can see the early Iron Age prologue to the urbanism of medieval and Renaissance times that ultimately shaped our modern world.



LUXURY PRODUCTS unearthed at Hascherkeller included glass beads (*a*), a bronze ring (*b*) and a bronze pin (*c*). Although the mold for casting the strips that were fashioned into finger rings (*d*) was found, no rings themselves were uncovered. The bronze pin and ring and the beads, like the graphite used to decorate the pottery, were probably traded for farm surplus.

Museum piece, circa 1987.

The clapstick may be taking an early retirement. This and other conventional tools of the filmmaking trade may find themselves relegated to crates and carted off to museums.

What will replace them? Kodak film with Datakode magnetic control surface! This radical advance in film manufacture, which unites chemical and magnetic imaging technologies, gives film the ability to "converse" with computers. This achievement could substantially reduce the time and costs associated with film postproduction.

The Datakode magnetic control surface is a thin, transparent layer of iron oxide particles (approximately 9 billion per square inch) coated across the entire back of the film during manufacture. Less than 8 microns thick, this transparent layer provides the means to record machine-readable information (100 bits of binary data in a single track on each frame of film), and makes possible a uniform frame-indexing code that can be used with both film and videotape. Best of all, it accomplishes this interface without altering the quality or characteristics of the final image.



In the not-too-distant future, the use of discs, video displays, time-code synchronization, and automated printing equipment will speed filmmakers through all the noncreative, repetitive steps associated with film postproduction.

The new Datakode magnetic surface is more than just another milestone for Kodak and a boon to filmmakers. It represents the most significant development in motion picture technology in the last 50 years. It's a victory for science. And for the chemists and researchers who met the challenge to add to film's capabilities.

For more information on the new Datakode surface, send your request to: Eastman Kodak Company, Department GBSA-7, 343 State Street, Rochester, NY 14650.

Kodak

Kodak. Where technology anticipates need.



© Eastman Kodak Company, 1983

EXTINCTION. IT'S FOREVER.

One quarter of all species of animals and plants on Earth may disappear in the next 30 years because of man's destruction of their habitat. The rate of extinctions is increasing enormously as forests are destroyed and other wild areas are lost. Organisms that evolved over hundreds of millions of years will be gone forever. The complex interdependence of all creatures, from the largest mammals to the smallest plants, is being shattered. It is a crisis with profound implications for the survival of all life. Unfortunately, little is being done to save our planet's natural heritage. Here are some warnings by leading scientists:

"The worst thing that can happen—will happen—in the 1980's is not energy depletion, economic collapse, limited nuclear war, or conquest by a totalitarian government. As terrible as these catastrophes would be for us, they can be repaired within a few generations. The one process ongoing in the 1980's that will take millions of years to correct is the loss of genetic and species diversity by the destruction of natural habitats. This is the folly our descendants are least likely to forgive us."

DR. EDWARD O. WILSON
Baird Professor of Science • Harvard University

"We are encroaching on nature, in the U.S. and around the world, at an unprecedented rate. A large proportion of the chemicals in use in our present-day civilization were 'invented' by nature, not by the chemist in the laboratory. An estimated 40% of all drug prescriptions in the U.S. contain as their chief ingredients compounds derived from plants. There is no end to the potential for discovery in nature, because we have only begun the chemical exploration of nature. Tragically, we are burning our library of priceless genetic treasures with our reckless destruction of species."

DR. THOMAS EISNER
Schurman Professor of Biology • Cornell University

"Few problems are less recognized but more important than the accelerating disappearance of Earth's biological resources. In pushing other species to extinction, humanity is busily sawing off the limb on which it is perched."

DR. PAUL EHRLICH
Bing Professor of Population Studies • Stanford University

"The extermination of a quarter of the plant species on Earth during the coming few decades could lose forever the medicines and food sources we need to cure the diseases and hunger that plague mankind."

DR. PETER RAVEN
Director • Missouri Botanical Garden

The Ark is sinking. We need the help of every concerned citizen to conserve the diversity of life on Earth. For information about how you can help, please write to us at:

Save Endangered Species

P.O. Box 50771 • Washington, D.C. 20004

Sponsored by:

- American Association of Zoological Parks and Aquariums
- American Cetacean Society
- American Humane Association
- American Institute of Biological Sciences
- Animal Protection Institute of America
- California Native Plant Society
- Center for Action on Endangered Species
- Center for Environmental Education
- Defenders of Wildlife
- Desert Fishes Council
- Environmental Defense Fund
- Florida Audubon Society
- Friends of the Earth
- Friends of the Sea Otter
- Fund for Animals
- Garden Club of America
- Greenpeace U.S.A.
- Humane Society of the United States
- International Council for Bird Preservation
- Massachusetts Audubon Society
- National Audubon Society
- National Parks and Conservation Association
- National Shell Fisheries Association
- Natural Resources Defense Council
- Ocean Research and Education Society
- Sierra Club
- Society for Animal Protective Legislation
- Wilderness Society
- World Wildlife Fund - U.S.

We offer executives



a much better position.



After those other airlines' San Francisco-to-Tokyo flights, you can be left feeling so bent out of shape that you half expect to be dumped off on the baggage carousel!

Not when you fly Philippine Airlines' new nonstops to Tokyo. We offer First Class passengers the only beds between San Francisco and Tokyo. Honest-to-goodness, full-length beds! 14 of them. Tucked quietly upstairs on our 747's where no one will disturb you—even during take-offs and landings.

So catch our new nonstops to Tokyo; catch our First Class, 5-star service from your seat downstairs; then catch all the sleep you need in your very own Skybed upstairs. And arrive in Tokyo feeling like a million.

Not like excess baggage.

Philippine Airlines now to Tokyo.

Amsterdam Bahrain Bandar Seri Begawan Bangkok Brisbane Dhahran Dubai Frankfurt Hong Kong Honolulu Jakarta Karachi Kota Kinabalu Kuala Lumpur London Los Angeles Manila Melbourne Paris Peking Rome San Francisco Seoul Singapore Sydney Taipei Tokyo Zurich

SCIENCE AND THE CITIZEN

Thinking Nuclear

Are U.S. nuclear weapons intended primarily to prevent a war with the U.S.S.R. or to fight one? That question, never far from the minds of professional military planners since the beginning of the nuclear-arms race, has come into increasing public prominence lately in the wake of the Reagan Administration's apparent efforts to shift the emphasis of U.S. policy from the former purpose to the latter. Since the early 1960's the prevailing strategic doctrine in Washington, if not in all corners of the military establishment, has been that the overriding objective of the nation's nuclear forces is to deter a preemptive nuclear attack by the U.S.S.R. on the U.S. or its allies. Indeed, notwithstanding the thousands of tactical nuclear weapons stationed with U.S. forces in Europe and the repeated assurances to Western European leaders that these weapons are meant in part to deter a conventional, or non-nuclear, invasion of their countries from the east, there is good reason to believe that for at least the past two decades no U.S. president would have ordered the use of nuclear weapons for any purpose other than to retaliate for a nuclear first strike by the U.S.S.R.

Support for this conclusion can be found in an extraordinary statement made recently by Robert S. McNamara, Secretary of Defense under presidents Kennedy and Johnson and one of the chief architects of the doctrine of nuclear deterrence. Writing in *Foreign Affairs,* McNamara reports that during the early 1960's "in long private conversations with successive Presidents—Kennedy and Johnson—I recommended, without qualification, that they never initiate, under any circumstances, the use of nuclear weapons. I believe they accepted my recommendation." Although careful to avoid suggesting that all U.S. presidents would behave in the same manner, McNamara quotes approvingly a revealing statement by another influential presidential adviser: Henry A. Kissinger. "The European allies," said Kissinger in a 1979 speech "should not keep asking us to multiply strategic assurances that we cannot possibly mean, or if we do mean, we should not want to execute because if we execute, we risk the destruction of civilization."

The currently favored alternative view, enunciated most notably by Secretary of Defense Caspar W. Weinberger, stresses instead the utility of nuclear weapons for waging a wide range of "limited" or "protracted" nuclear wars and the concomitant undesirability of a "no first use" policy. Although Weinberger and other Administration officials have tended to mute their comments on this issue somewhat since the early days of President Reagan's term in office, the material evidence of their planned restructuring of the nation's military priorities is clearly discernible in the number, size and variety of the nuclear-weapons programs included in the Administration's past and future military-budget requests.

According to a recent study by budget analysts at the Center for Defense Information, spending on nuclear weapons has more than doubled since President Reagan took office three years ago. Over the next six years, they report, the share of the military budget devoted to preparations for nuclear war is projected to continue its rapid growth, accounting for some $450 billion, or 22 percent of the total military outlay for that period. By 1992, the report says, some 17,000 new nuclear weapons will have been made in the U.S.—an average of five per day for the next decade. Assuming the retirement of approximately 11,000 older weapons over the same period, the analysts remark, "the U.S. nuclear stockpile could grow by an estimated 6,000 bombs and warheads."

Included in the Administration's military-budget projections for the next 10 years are funds for 15 different types of nuclear warheads, aerial bombs and artillery shells. Nine of these new explosive devices are already under construction and work on several others is scheduled to begin soon. In general, the Center for Defense Information analysts observe, "the delivery systems these weapons will be fitted to are more sophisticated, more accurate and more destructive than those they are replacing." More to the point, they add, "many of the weapons being sought and the reasons given for their necessity fit a nuclear-war-fighting strategy."

Not surprisingly, the increasing reliance of the nation's military services on nuclear weapons is also reflected in a progressive "nuclearization" of training exercises and other war preparations. The effect of this trend is evident in testimony by Admiral Powell Carter before the Senate Armed Services Committee. "We have revamped the training command in order to teach our people how to think nuclear," he is reported to have said. "We are extending to where all our war exercises are into a nuclear phase."

It is precisely such thinking on the part of many of today's military and political leaders that former high officials such as McNamara evidently are beginning to find unsettling. "Having spent seven years as Secretary of Defense dealing with the problems unleashed by the initial nuclear chain reaction 40 years ago," he writes, "I do not believe we can avoid serious and unacceptable risk of nuclear war until we recognize—and until we base all our military plans, defense budgets, weapon deployments and arms negotiations on the recognition—that *nuclear weapons serve no military purpose whatsoever. They are totally useless—except only to deter one's opponent from using them.*"

Whirl Pool

The discovery of the three long-sought intermediate vector bosons at the European Laboratory for Particle Physics (CERN) this year has quickened the efforts of particle physicists around the world who seek a better understanding of the unified nature of the weak force and the electromagnetic force. The strong experimental evidence favoring the existence of the W^+, the W^- and the Z^0 particles confirms one of the main predictions of the electroweak theory; the theory regards the weak force and the electromagnetic force as two aspects of a single underlying phenomenon. The success of the theory suggests to many physicists that its predictions about interactions of still higher energy may well be true; confidence in the theory now seems high enough to justify the considerable risk in time, effort and money that will be needed to continue its verification. In the next 10 to 15 years a substantial part of the resources of other major laboratories will be devoted to testing and probing the predictions of the electroweak theory and, not incidentally, to seizing the mantle of leadership from CERN.

In the U.S. the results at CERN have emboldened the physics community to begin serious work on the design of a machine nicknamed the Desertron, so called for one of its likely sites. The Desertron, more formally called the Superconducting Super Collider, would accelerate two beams of protons in opposite directions around a track perhaps 100 kilometers in circumference. Protons in each of the two beams would reach an energy of about 10 TeV (10 trillion electron volts) before colliding, thereby liberating twice that energy in the collision. The net energy available for the materialization of new particles would be on the order of 1 TeV, or roughly 40 times the energy that can be extracted from the present machine at CERN. Many physicists believe that at such energies a new particle called the Higgs particle, required by the structure of the electroweak theory, could be created.

The detection of a Higgs particle would be a discovery of the first rank,



A partridge in a pear tree



Two turtledoves



Three French hens



Four calling birds



Five gold rings



Six geese a-laying



Seven swans a-swimming



Eight maids a-milking



Nine ladies dancing



Ten lords a-leaping



Eleven pipers piping



Twelve drummers drumming

What people gave before there was Chivas Regal.

To send a gift, dial 1-800-528-6148 12 YEARS OLD WORLDWIDE · BLENDED SCOTCH WHISKY · 86 PROOF · © 1982 GENERAL WINE & SPIRITS CO., NEW YORK, N.Y.

The Texas Instruments
Because y(x)=(1/k) co



Home Computer.
sh [k(x+a)]+b.

Whether it's beginning math or this catenary equation, the TI-99/4A Home Computer gives you real power to solve real problems.

A true home computer should be user friendly to every family member, novice or professional. It takes power to do it.

That's what makes the TI-99/4A Home Computer the one to choose. It's a powerful package designed around our 16-bit TMS 9900 microprocessor. And it contains a lot of fully usable memory – 16K RAM and 26K ROM, plus up to 38K more ROM in each solid state cartridge.



The problem: A 100-foot-long cable, whose weight is 0.67 pounds/foot (corrected for buoyancy) is used to secure a mooring buoy to an anchor in 30 feet of water in a flat-bottomed lake. Determine the horizontal and vertical force components at the anchor and buoy when the buoy is displaced horizontally 92 feet from the anchor. If any of the cable lies on the bottom of the lake, find the distance from the anchor at which it breaks away from the bottom. When the cable is unconstrained (i.e., not touching bottom), the equation of the catenary is $y(x) = (1/k)\cosh[k(x+a)]+b$. See the computer screen below for the solution.

With TI's BASIC system built into ROM, it doesn't use up Random Access Memory (RAM). Others, who load their operating systems into RAM, can use up half their memory before you even get started.

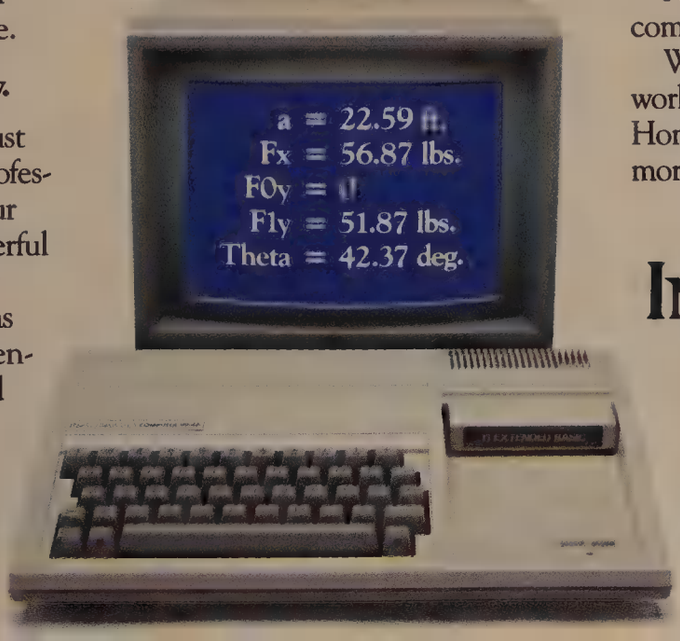
The result is that the TI Home Computer performs highly sophisticated computing tasks as easily as it teaches third grade math. For the money, no other computer is better suited for technical or scientific use.

Greater arithmetic accuracy.

Our unique 16-bit processor is just like the one found in expensive professional computers. Coupled with our rich TI BASIC, or still more powerful TI Extended BASIC, it makes our scientific programming capability as good as that provided by many scientific languages for mainframes, and better than some.

We give you 13-digit accuracy, and number range spanning $1x10^{\pm 127}$. All other home computers, using 8-bit processors, limit you to 6 to 9 digits, and only $1x10^{\pm 38}$ range. It's a big difference. And it means you can solve problems on the TI Home Computer which aren't feasible on the others.

TI BASIC also gives you more powerful commands and better flexibility for naming variables, since it recognizes up to 15 character descriptive names. Some others recognize only the first two significant characters in names, which means they can't tell the difference between "TAX" and "TARIFFS." So you have to resort to confusing acronyms.



The TI Home Computer is fluent in seven languages.

You can choose languages for all programming levels, from built-in TI BASIC to TI Extended BASIC, TI LOGO II, UCSD Pascal,* TI PILOT, TI FORTH and our powerful TMS 9900 Assembly Language.

You can also write useful non-scientific programs with the TI Home Computer. It's the only one in its price range that lets you combine 16-color graphics, single-command animation, sound effects, music, and speech with unlimited vocabulary into the same TI BASIC program.

And only TI's home computer comes with a 1-year warranty.

Whatever kind of problem you're working on, work it out with the TI Home Computer. You can't find more computer for the money.

TEXAS INSTRUMENTS

Creating useful products and services for you.

*UCSD Pascal is a trademark of the Regents of the University of California.

Copyright © 1983 Texas Instruments

and the possibility that the Desertron could find it has mobilized the American physics community behind the machine. In June, citing the need to consolidate funds and talent, a subpanel of the High Energy Physics Advisory Panel (HEPAP) of the Department of Energy voted 10 to seven to abandon construction of the half-finished Colliding Beam Accelerator at the Brookhaven National Laboratory in favor of the Desertron. Soon after the vote the members of HEPAP unanimously agreed to make the Desertron their top priority.

In October the Department of Energy redirected $18 million originally designated for the Colliding Beam Accelerator to a feasibility study for an ultrahigh-energy accelerator. Research and development over the next three to four years is expected to cost $150 to $200 million. The total cost of the machine is estimated to be at least $2 billion, and the length of time needed to design and build it would probably be 10 to 12 years. Nevertheless, many physicists are convinced that whether or not the Higgs particle is found, the Desertron would reach energy regions of great importance to physics and must be built if the most talented young physicists are to be attracted to the U.S. program.

There has been a major technological advance that makes it feasible to contemplate the Desertron, namely the development of reliable superconducting magnets. Such magnets can generate enormous magnetic fields for steering and confining a beam of particles, and they are far less costly to operate than nonsuperconducting electromagnets because superconducting materials have no resistance to the flow of electric current. Superconducting magnets were to have been installed at the aborted Brookhaven accelerator, and the difficulties encountered at Brookhaven in their manufacture caused the delays in construction that finally led to the cancellation of the project. The problems with the magnets have now been solved. According to Nicholas P. Samios, director of Brookhaven, the Desertron might be a version of the Brookhaven accelerator scaled up by a factor of 50.

Superconducting magnets have been installed at the Tevatron I, the main accelerator ring at the Fermi National Accelerator Laboratory (Fermilab); the ring will eventually accommodate countercirculating beams of protons and antiprotons. According to Leon M. Lederman, director of Fermilab, a single beam of protons reached a world-record energy of 700 GeV (700 billion electron volts) in August, and work is now focused on raising the luminosity, or intensity, of the beam and on extracting it for bombardment of a fixed target. The Tevatron I is to be tested and operated as a fixed-target machine until the summer of 1985; in other words, the high-energy protons will be directed at matter that is at rest instead of against particles in a countercirculating beam. Thereafter the machine is to be run as a colliding-beam accelerator. The energy released by the collision of protons and antiprotons is expected to reach 2 TeV.

The Tevatron I is the major competitor of the Super Proton Synchrotron, the proton-antiproton collider at CERN that was responsible for the discovery of the W and Z particles. Indeed, once the Tevatron I begins to operate at full power as a colliding-beam machine, which Lederman expects by the summer of 1986, it will surpass the energy and luminosity capabilities of the CERN accelerator.

Accordingly there is much interest at CERN in carrying out as many new experiments as possible before the Tevatron is ready. The energy capability of the CERN machine will be raised slightly to 310 GeV for each beam, and a major effort will be made to increase the luminosity of the beam. The current plan calls for the construction of a new collector for antiprotons, which would cost about $20 million and would increase the luminosity of the antiproton beam by a factor of 10. These improvements would guarantee a much higher production rate of the W and Z particles than is now possible.

In September ground-breaking ceremonies were held for another machine at CERN called the Large Electron-Positron collider (LEP); the LEP should generate about 10,000 vector bosons per day when it begins operation in 1988. The main tunnel in the LEP collider will be 27 kilometers long, and in its first phase of operation the machine will enable electrons and positrons to collide at energies of up to 60 GeV per particle. Eventually that energy will be raised to 100 GeV. Although the energy generated by the collisions is far less than that generated by the heavier protons and antiprotons, a large fraction of the electron-positron energy becomes available for the creation of new particles. (Because the proton and the antiproton are made up of quarks and gluons, the energy supplied to the proton and the antiproton must be divided among several elementary particles.) There is no firm theoretical prediction for the mass of the Higgs particle, and so the LEP could well find the particle before the Desertron is built. Moreover, there is discussion at CERN that the LEP tunnel could later be converted into a proton ring, where energies on the order of 8 TeV could be achieved.

Until the LEP is built much of the physics generated by electron-positron interactions will be carried on at two U.S. laboratories. The HEPAP subpanel has also recommended the upgrading of the Cornell Electron Storage Ring and the rapid construction of the Stanford Linear Collider. The latter, to be built at the Stanford Linear Accelerator Center (SLAC), will accelerate electrons and positrons along the track of the existing three-kilometer linear accelerator to 50 GeV. The two beams are then to be split by an electromagnet and sent in opposite directions halfway around a ring, where they will collide. Construction is now under way on the $113 million project and should be completed by the end of 1986. The machine will generate copious numbers of W and Z particles, and it will also test the linear accelerator as a cheaper alternative to the well-established storage ring.

At the Deutsches Elektronen Synchrotron (DESY) in Hamburg construction is to begin early next year on a device called the Hadron-Electron Ring Accelerator (HERA). The name points to the distinctive feature of the accelerator: it will be the only machine in which a hadron, such as the proton, is made to collide with a lepton, such as the electron. The design will enable physicists to closely examine the effects of the weak nuclear force without having to account for large effects of the electromagnetic force or the strong nuclear force, which can obscure the data collected from other machines. One major goal of the HERA machine is to penetrate the field surrounding the quark to within 10^{-17} centimeter. If the quark has a finite size, the size may become evident at that scale. It is left to the more powerful Desertron to determine the shape and structure of the quark, if they exist. The HERA is expected to go into experimental service in 1990.

At the 12th International Conference on High Energy Accelerators held at Fermilab in August physicists heard several reports on the development of accelerators in China, in the U.S.S.R. and in Japan. Outside Beijing a 3-GeV electron-positron collider is being constructed that should go into operation in 1987. The Russian laboratory at Serpukhov is building a 400-GeV proton accelerator called UNK, scheduled for completion in 1988. The accelerator is to be upgraded in energy to 3 TeV in the 1990's, and a second beam of 3-TeV protons is eventually to collide with the first. The Japanese project Tristan, started in 1981, is to be finished by 1986. A beam of 30-GeV electrons will collide with a beam of 30-GeV positrons.

Unruly Child

An El Niño is an appearance in the central and eastern equatorial Pacific of anomalously warm water. El Niños occur at irregular intervals of three to 10 years in conjunction with a large-scale fluctuation in atmospheric pressures known as the Southern Oscillation. The consequences of these coupled phenomena are often severe, ranging

Ford Escort Diesel: Better mileage than this leading import.



We didn't believe it at first, either.

But EPA testing figures established it. Our new Escort Diesel is rated approximately four miles per gallon higher than a Honda 750.

Just take a look at our numbers:

46 EPA EST. MPG.

68 EST. HWY.

And because this diesel is an Escort, there's a lot more to talk about than great economy.

Like the fact that Escort's the best-selling car in the world.**

Or that it comes with more total passenger room and more total cargo room than a Honda Accord.†

More standard features than a Toyota Tercel.††

There's even a fully-independent suspension system for a smoother ride than a Nissan Sentra.

All of which means Ford Escort not only gives you a big advantage over that motorcycle pictured above.

It also beats more than its share of cars.

THE BEST-BUILT AMERICAN CARS.

When we say "Quality is Job 1," we are talking about more than a commitment. We are talking about results. An independent survey concluded Ford makes the best-built American cars. The survey measured owner-reported problems during the first three months of ownership of 1983 cars designed and built in the U.S.

And that commitment continues in 1984.

* For comparison. Honda 750 mileage is obtained from EPA emissions testing and is *not* an official rating. Your mileage may vary depending on speed, trip length, weather. Actual highway mileage lower. Escort Diesel mileage applicable to sedans with FS engine and without power steering and A/C. Not available in California.

** Sales estimates based on world-wide production figures.

† Based on EPA Interior Volume Index.

††Escort GL (shown) compared to Toyota Tercel 3-door deluxe liftback.

Get it together – Buckle up.

Have You Driven A Ford... Lately?

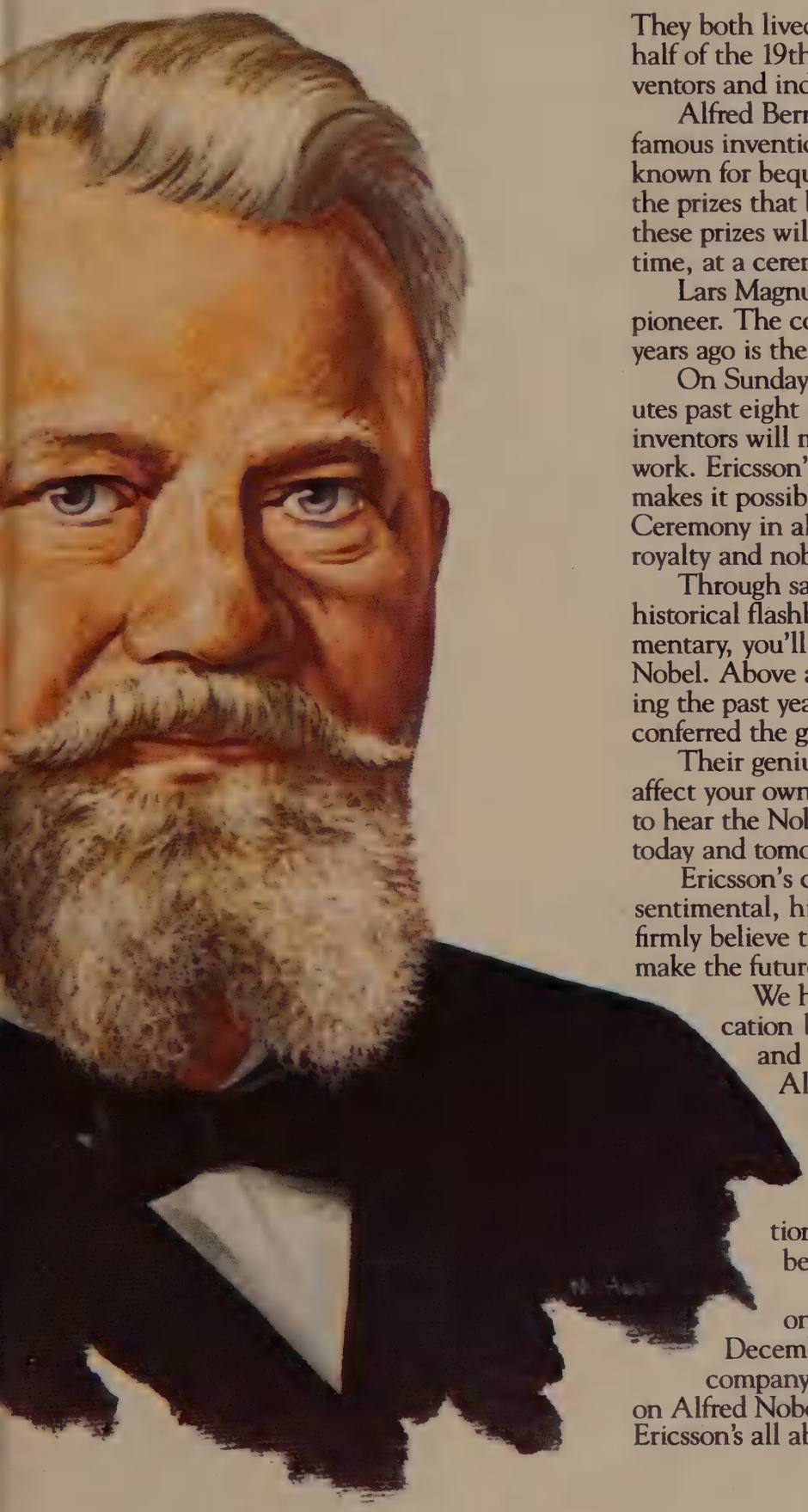
Ford





A. B. Nobel, 1833–1896

ONE MAN'S NAME IS ON EVERYBODY'S LIPS. THE OTHER ONE'S SOON WILL BE.



They both lived in Stockholm during the second half of the 19th century. Both were prominent inventors and industrialists.

Alfred Bernhard Nobel was a chemist. His most famous invention was dynamite. But he is best known for bequeathing his vast fortune to institute the prizes that bear his name. On December 10, these prizes will be awarded again, for the 82nd time, at a ceremony in Stockholm's City Hall.

Lars Magnus Ericsson was a telecommunications pioneer. The company he started over a hundred years ago is the world's fourth largest in its field.

On Sunday, December 11, at exactly five minutes past eight in the evening (EST), these two inventors will meet again—on WTBS Cable Network. Ericsson's role as a Nobel Fund contributor makes it possible for you to view the Nobel Awards Ceremony in all its pomp and circumstance, with royalty and nobility in attendance.

Through satellite coverage, interspersed with historical flashbacks and up-to-the-minute commentary, you'll get to know the enigmatic Alfred Nobel. Above all, you'll meet the people who, during the past year, to quote Nobel himself, "have conferred the greatest benefit on humankind."

Their genius and their insights will profoundly affect your own future. Don't miss this opportunity to hear the Nobel Laureates' views on the issues of today and tomorrow.

Ericsson's contribution to this event is not for sentimental, historical reasons, but because we firmly believe that science and technology can help make the future better for all of us.

We have been helping to make communication between individuals, organizations and nations easier ever since the days of Alfred Nobel. Today, we're known as the pioneer in the digitalization of the global telephone network.

And the more we innovate in areas like data processing, information systems and cellular radio, the better you, too, will get to know us.

Watch the Nobel Prize Ceremony on WTBS Cable Network on Sunday, December 11. And write to us, on your company letterhead, for a free 12-page booklet on Alfred Nobel and his work—and a little on what Ericsson's all about.

L.M. Ericsson, 1846–1926

ERICSSON

ERICSSON, INC., GREENWICH OFFICE PARK 3, GREENWICH, CT 06836-3110

from droughts in Australia and South Africa to devastating rains and sharply reduced fish catches in Ecuador and Peru. The 1982–83 El Niño has been the strongest in this century: typically, sea-surface temperatures off the South American coast rise by two to three degrees Celsius during an El Niño, but the anomaly reached a peak of seven degrees in June. The anomaly was also anomalous in that it began later in the year than previous events and did not announce itself with the usual atmospheric signal.

During a Southern Oscillation sea-level air pressure subsides in the high-pressure system of the southeastern Pacific and rises in the low-pressure area normally over Indonesia and northern Australia. The result is a decrease in the pressure gradient that drives the easterly trade winds along the Equator. Under normal conditions these winds and the westward equatorial current pile up very warm water (29 degrees C.) in the western Pacific. When the trade winds slacken, the warm water can slosh back eastward, leading to anomalously warm sea-surface temperatures in the central Pacific. The wind shift also appears to trigger a series of subsurface waves, called Kelvin waves, that propagate along the Equator as vertical displacements of the thermocline (the boundary between the warm upper water of the ocean and the cold water below it). The Kelvin waves depress the thermocline and thereby thicken the warm surface layer.

It is the latter phenomenon on which oceanographers now rely to account for the appearance, sometimes with disastrous ecological and economic consequences, of unusually warm water off the South American coast. Normally the southeast trade winds blowing along the coast push surface water out to sea, allowing cold water from below the thermocline to rise to the surface. After the Kelvin waves hit the coast, however, the sea level rises and the warm surface layer gets so thick that the water welling up is warm, not cold. Moreover, it does not contain the rich supply of nutrient salts that the cold water normally brings to the surface and that phytoplankton require for photosynthesis. The resulting decline in plankton levels is felt all the way up the food chain, ending with the sea birds and fishermen who work the usually bountiful sea.

Previous El Niños had been preceded by a buildup of the easterly trade winds. Following the drop in the winds in the fall, warm water tended to appear first off South America just after Christmas, in January or February (hence the term El Niño, literally "the little one" but referring to the Christ child). The peak of the anomaly came in May or June. This event was different: the trades never peaked before slackening, and they did not slacken until April, 1982. Warming of the sea surface began in the central Pacific in the summer and in the eastern Pacific in September.

Investigators had come to regard a buildup of the trade winds as being essential for the piling up in the western Pacific of warm water that would then slosh back eastward during the El Niño. They are now reconsidering this concept. The odd timing of this year's event is also a puzzle, but according to Mark A. Cane of the Massachusetts Institute of Technology it need not be. Cane suggests that one of the first things to happen in an El Niño is a movement of the massive Indonesian low-pressure system eastward along the Equator toward the central Pacific. In previous events such movement has tended to come in the fall, but there is no reason it should not come in the spring, as it did last year; in both seasons the low-pressure zone is less powerfully centered, more unstable and therefore more mobile.

Why the Indonesian low should move in some years and not in others is not known. Once it starts moving, however, a feedback mechanism could keep it going and eventually produce an El Niño, according to Eugene Rasmusson of the National Meteorological Center. Like other lows, the Indonesian low is a convergence zone, where air is heated and rises and is replaced by winds flowing into the zone. Convergence zones are generally over areas where the surface is warmest and evaporation and upward motion of the air are therefore greatest.

When the Indonesian low is in its normal position over Indonesia, the inflow into the system over the western Pacific is provided by the easterly trade winds. As the low moves eastward toward the International Dateline in the middle of the Pacific, however, it draws air in behind it from the west, creating anomalous westerlies that damp the trade winds. Now the warm pool of water in the western Pacific can begin to move eastward, taking with it an area of heavy rainfall. The precipitation warms the atmosphere, intensifies the westerlies (in 1982–83 the trade winds actually reversed direction in the western Pacific) and draws the low-pressure zone farther east. Meanwhile the signal that the winds have shifted in the west is propagating along the Equator toward South America, at a speed of several hundred kilometers per day, as a series of Kelvin waves.

How is this signal transmitted to the mid-latitudes of the Northern Hemisphere? When an equatorial El Niño occurs, almost invariably a mid-latitude El Niño produces smaller sea-surface temperature anomalies (two to three degrees C. in 1982–83) off the North American coast. Although the present equatorial El Niño appeared to be fading away this past fall, the mid-latitude anomaly, which has reached as far north as British Columbia and Alaska, is if anything showing signs of getting stronger. That the two events are connected seems obvious, but the nature of the connection is not.

According to one hypothesis, the equatorial Kelvin waves split in two when they hit South America and travel north and south along the coast, carrying warm water into the mid-latitudes and depressing the thermocline as they go. James J. Simpson of the Scripps Institution of Oceanography reports, however, that the warm water reaching the California coast during the present El Niño contains too little salt and too much dissolved oxygen for it to be coming from the south. Detailed measurements made by Simpson and his colleagues along a line extending southwest from San Diego indicate instead that the warm water comes primarily from the west and is being driven onshore by the same anomalous southwesterly winds that battered California with storms all last winter. These results suggest a primarily atmospheric rather than oceanic "teleconnection" between the equatorial and mid-latitude El Niños.

Arthur Douglas of Creighton University offers a possible scenario for such a teleconnection. In normal years the low-pressure zone near the Equator where the northeast and southeast trade winds meet, called the Intertropical Convergence Zone (ITCZ), migrates north in May or June. During an El Niño, for reasons that are not entirely clear, the ITCZ stays south. As a result the upper-level high engendered in the subtropics by excessive air rising from the ITCZ stayed over Mexico in 1982, producing hot, dry weather that allowed the Pacific to warm up as far as 700 miles offshore.

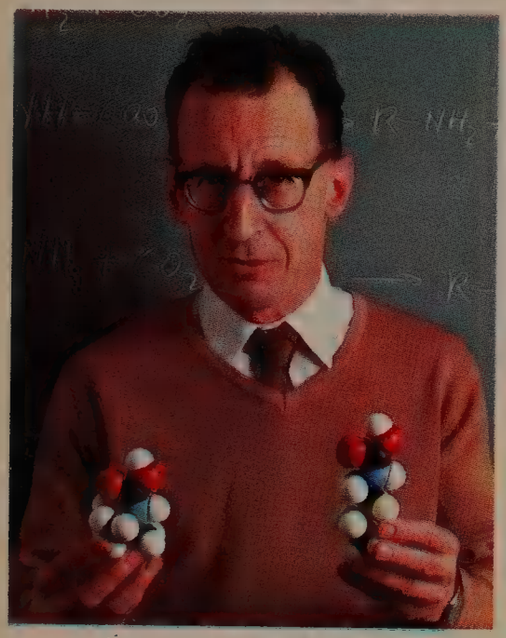
Heat energy from this pool of warm water then intensified the southwest winds flowing into the upper-level low-pressure trough that often lodges over California in September. These winds drove warm water onto the coast and prevented the normal upwelling of cool water. By winter the warm anomaly had spread to British Columbia as the onshore winds associated with a low-pressure system over the Gulf of Alaska intensified. The forces driving this last process are also not well understood, but they are assumed to be related to the large temperature contrast between the wide band of El Niño water at the Equator and a pool of anomalously cold water observed this year in the central Pacific.

Douglas thinks a similar sequence of events has taken place in previous El Niños, but with differences that account for the vastly different effects the events have had on North American weather. During the El Niño of 1976–77 California had a drought and the winter in the U.S. was unusually cold. The current


Savor the
sense of Rémy.
REMY MARTIN
Rémy Martin
V.S.O.P.
COGNAC
Rémy
REMY MARTIN V.S.O.P. COGNAC. SINCE 1724.
Imported by Rémy Martin Amerique, Inc., N.Y.
80 Proof

How Exxon developed can double the productivity

Guido Sartori's work on hindered amines may impact an entire industry.



Removing impurities such as carbon dioxide and hydrogen sulfide from natural, refinery, and synthesis gases is an expensive, energy-consuming process.

But at Exxon Research and Engineering Company a new chemistry discovery, and cross functional teamwork, have led to the development of a new technology—one that significantly decreases the cost and increases the capacity of commercial gas treating processes.

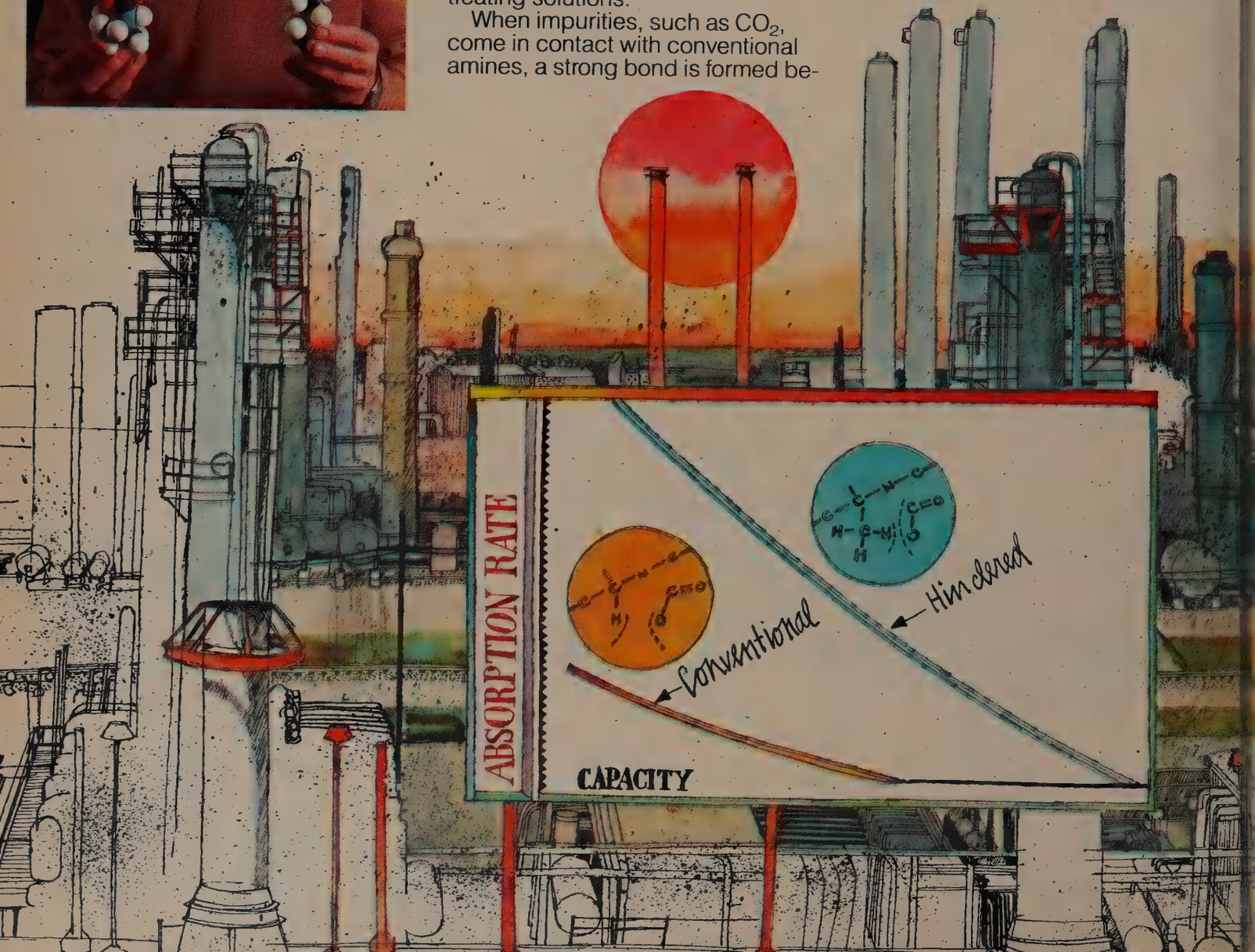
Research Led to a Discovery

Guido Sartori, a chemist in Exxon Research and Engineering Company, had been conducting research on amines—organic nitrogen-containing molecules—to increase both the absorption rate and capacity of gas treating solutions.

When impurities, such as CO_2, come in contact with conventional amines, a strong bond is formed between the CO_2 and the nitrogen ato of the amine. This strong bond ties a disproportionate amount of usefu amine. Sartori theorized that both th absorption rate and capacity of the amine would be improved if the bon at the nitrogen site could be weakened. Continuing research revealed the advantages of a whole new clas of amines, which he called hindere amines.

Observing Molecular Behavior

Sartori and others began a comprehensive evaluation of the discovery, utilizing the company's advance analytical capabilities. To understan the behavior of hindered amines, an to monitor reactions, Sartori employ the results of carbon-13 nuclear ma netic resonance spectroscopy, a



new molecules that of gas treating plants.

state-of-the-art technique not previously used for this purpose.

Further research confirmed the hindered amines' capability to substantially increase the rate and capacity of carbon dioxide absorption through the formation of low stability bonds. Low stability was achieved by placing a bulky substituent next to the nitrogen sites, thereby hindering bond formation with CO_2. Building on this new understanding, he synthesized new molecules to meet the performance requirements for specific applications.

Integrated Innovation

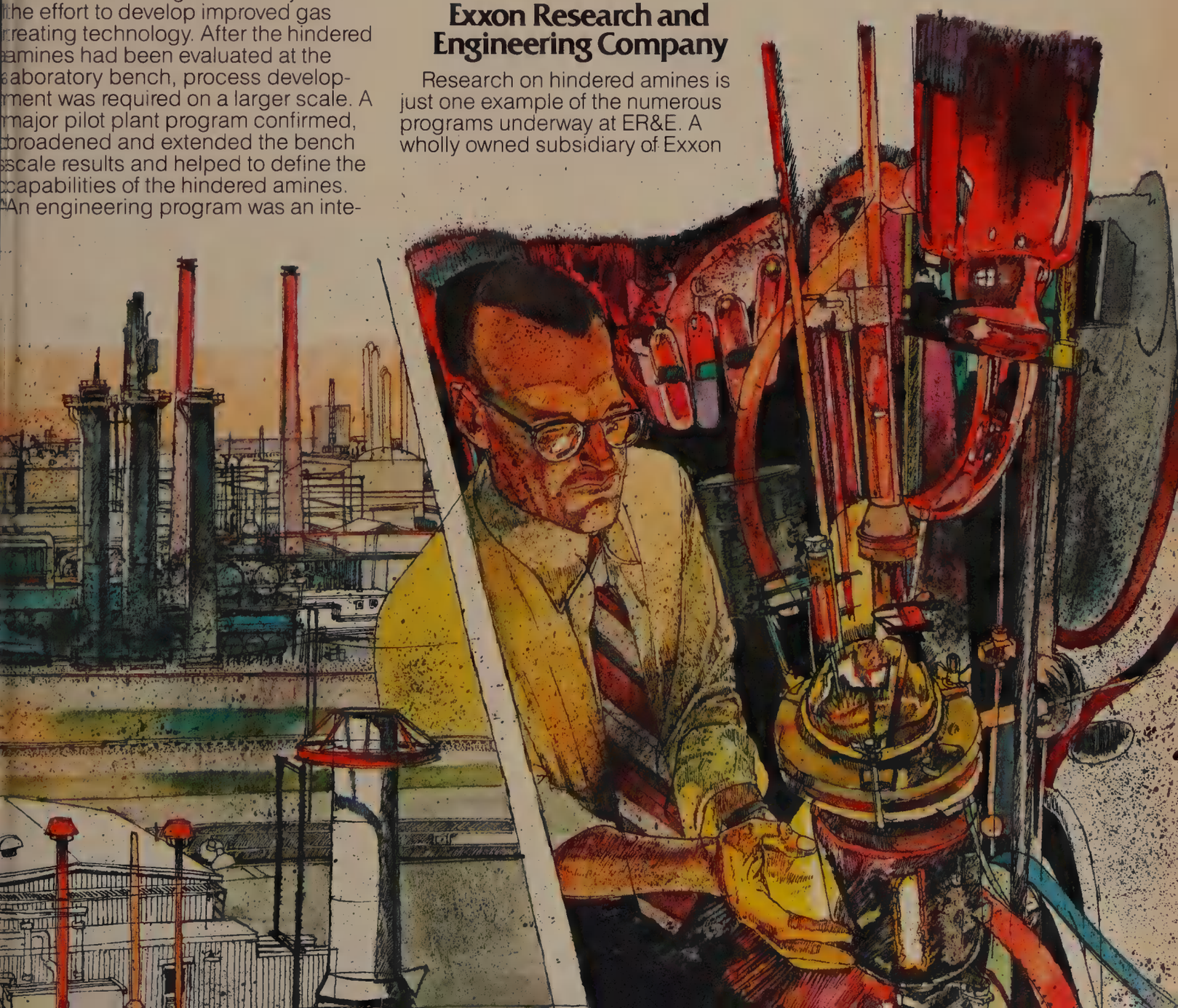
Other Exxon organizations joined the effort to develop improved gas treating technology. After the hindered amines had been evaluated at the laboratory bench, process development was required on a larger scale. A major pilot plant program confirmed, broadened and extended the bench scale results and helped to define the capabilities of the hindered amines. An engineering program was an integral part of the research and development required to convert these laboratory discoveries into commercially feasible technologies. Capacity increases of 50% have been achieved commercially using this technology with no added facilities.

Through integrated innovation—the combined efforts of the company's basic research, process development, and engineering staffs—hindered amine technologies advanced from scientific discovery through commercial use in less than three years. Further research has enabled ER&E to identify or synthesize other practical hindered amines.

Exxon Research and Engineering Company

Research on hindered amines is just one example of the numerous programs underway at ER&E. A wholly owned subsidiary of Exxon Corporation, ER&E employs some 2,000 scientists and engineers working on petroleum products and processing, synthetic fuels, pioneering science and the engineering required to develop and apply new technology in the manufacture of fuels and other products. For more information on Exxon's hindered amine technology or ER&E, write Dr. E. E. David, President, Exxon Research and Engineering Company, Room 705, P.O. Box 101, Florham Park, New Jersey 07932.

EXXON



event produced downpours in California and may account for the mild winter of 1982–83 in most of the U.S.

Earlier Americans

When did men out of Asia first enter the New World? One position on the question, held by many conservative-minded archaeologists, is doubtful of any date earlier than some 12,000 years ago. Another position, held by a few venturesome workers, rests on a single California site: Calico, an ancient alluvial fan in the Mojave Desert. Geomorphologists consider the fan to be from one to two million years old, but proponents of Calico as an early-man site suggest that fractured stones from the lower part of the fan are manmade artifacts some 200,000 years old.

Richard S. MacNeish of Boston University reviews the question in the light of archaeological findings, including his own near Ayacucho in Peru. Writing in *The L. S. B. Leakey Foundation News,* he addresses his main argument to his conservative colleagues: the proposal that man pushed eastward across the Bering Strait as recently as 12,000 years ago must be dismissed if earlier evidence of immigration exists in South America.

MacNeish points out that just such dates have been established by carbon-14 analyses of material from the lower levels of one of his Ayacucho sites, Pikimachay Cave. Two separate analyses from the cave's lowest man-occupied stratum yielded respective dates of 19,600 and 20,200 B.C. The dates of three overlying strata are 16,100, 14,750 and 14,100 B.C. MacNeish goes on to list other South American findings: (1) at Los Toldos in southern Argentina stone tools have been found in levels older than 10,650 $\pm$ 600 B.C., (2) at the Piaui Caves in Brazil flake tools older than 16,650 $\pm$ 600 B.C. have been found in lower levels, (3) at Monte Verde in Chile tools have been found with mastodon bones dating to about 12,000 B.C. and (4) at Taima-Taima in Venezuela a projectile point has been found embedded in a mastodon pelvis; although its carbon-14 age has not been published, the date is estimated at 14,000 B.C.

Turning to North America, MacNeish points to Valsequillo and Tlapacoya in Mexico. At Valsequillo tools in association with the bones of extinct animals are estimated to range in date from 20,000 to nearly 40,000 B.C. At Tlapacoya hearth charcoal with a carbon-14 age of 23,150 $\pm$ 500 years before the present is associated with both artifacts and the bones of extinct animals.

Commenting on the Los Toldos site in Argentina, the one with a date closest to 12,000 years ago, MacNeish remarks: "No matter how fast man ran the length of [South America] he had to come across Bering Strait well before 12,000 years ago in order to drop these tools at this date." He goes on to report a number of recent defections from the conservative ranks, perhaps most notably that of C. Vance Haynes of the University of Arizona. Haynes now places the date of the first New World immigration at more than 30,000 years ago.

As for the Calico site enthusiasts, either of two factors may weaken their position. If, as geomorphologists assert, the alluvium was deposited more than a million years ago, no archaeologist would suggest that the deposit contains human handiwork. If, on the other hand, the deposit is only a fraction as old, it still remains to be established whether the Calico objects are crude stone tools worked by man or stones worked by natural agencies. At present arguments are offered on both sides of the question.

Forests of the Night

The hallmark of a quasar is a set of spectral lines prodigiously out of position. The Lyman lines are the best example. They consist of the photons (the quanta of light) emitted by electrons in hydrogen atoms when they "fall" from excited states to the ground state: their state of minimal energy. In the laboratory the Lyman lines are found in the ultraviolet; Lyman alpha, the most prominent line in the set, has a wavelength of 1,216 angstrom units. Yet in the light from a typical quasar the wavelengths of the Lyman lines are several times longer: they lie in the visible part of the spectrum. In the most widely accepted hypothesis quasars are taken to be "engines" at the core of newly formed galaxies extremely distant in space and in time, and the systematic displacement of the lines in the light each one emits is taken to be a red shift: an increase in wavelength indicative of the quasar's recession from the earth as part of the general expansion of the universe.

In an effort undertaken by Wallace L. W. Sargent of the California Institute of Technology and Alec Boksenberg of University College London some Lyman-alpha lines in the light emitted by quasars are significant in quite a different way. They are not emission lines. They are absorption lines, that is, they are dips, not peaks, in the spectrum. They are caused when the electrons in hydrogen atoms pluck light from a quasar's broadband radiation and are excited from the ground state into a state of greater energy. In the light from a single quasar these Lyman-alpha absorptions can number more than 100. Each has a different wavelength longer than 1,216 angstroms but shorter than the red-shifted Lyman-alpha emission of the quasar itself. Sargent and Boksenberg propose they represent the absorption of quasar light by clouds of hydrogen at different positions along the line of sight to the quasar. Each cloud would be receding from the earth at some velocity less than that of the quasar, and so its hydrogen atoms would absorb the quasar's light at a distinctive red-shifted value. The resulting "forest" of Lyman-alpha absorptions means that the light of the quasars can serve as a probe of the hydrogen clouds in intergalactic space.

Sargent and Boksenberg approach the Lyman-alpha forest by removing everything else. Some of the absorption lines they encounter in the light of a typical quasar are relatively prominent and are red-shifted from their laboratory wavelengths by only a small amount. Such lines are attributed to the absorption of quasar light by elements such as carbon, oxygen, silicon, iron and magnesium in the gas and dust that permeate our own galaxy. Other lines are less prominent; they turn out to be absorption lines due to hydrogen accompanied by absorption lines due to heavier elements, all moved from their laboratory wavelengths by an identical red shift. They are attributed to the absorption of quasar light by galaxies other than our own along the line of sight to the quasar. Still other lines are hydrogen lines at a multitude of red shifts. They make up the Lyman-alpha forest. They prove not to be accompanied by lines due to heavier elements; thus Sargent and Boksenberg calculate that the abundance of heavy elements relative to the abundance of the hydrogen can be no greater than a thousandth the abundance of heavy elements in the sun. Evidently the clouds consist of primordial hydrogen that has never condensed into stars.

The properties of the clouds are beginning to emerge. For one thing, Sargent and Boksenberg note that the light from quasars that have small angular separations in the sky can reveal the size of the clouds. They have already found that the Lyman-alpha forests in the spectra of 10-billion-year-old quasars one minute of arc apart are different from one another. They conclude that the clouds are probably smaller than .5 megaparsec, roughly the size of a cluster of galaxies. More generally, Sargent and Boksenberg report that the lines in a Lyman-alpha forest show no grouping at particular red shifts; hence the clouds do not group together in space. In this they differ from galaxies. On the other hand, the lines are more numerous at greater red shifts; hence the ratio of clouds to galaxies increases with increasing distance in space and in time. That is, the clouds were more numerous in the early universe. What made their numbers decrease? Not gravitational collapse, or else stars would have formed. Sargent and Boksenberg suspect the clouds are held together by the pressure of the surrounding intergalactic medium, a hydrogen gas whose atoms are ionized by the photons streaming from quasars.

ductivity Network.



Presenting your broadest choice of engineering workstations anywhere. From a low-cost personal computer to a powerful 32-bit system.

No doubt about it: when you put a computer on every engineer's desk, it helps them be more productive.

In fact, at Hewlett-Packard, the workstation approach has worked so well for our own engineers that we've taken it to an entirely new dimension.

It's called the Engineering Productivity Network. As you can see, it includes the broadest range of workstations available anywhere. From the recently-introduced 32-bit HP 9000 desktop mainframe to our personal computer designed just for engineers.

With such a wide range to choose from, you can match the computing power to the individual task. And give every engineer the tool it takes to do the best possible job.

To help you tailor every workstation to a task even more closely, we have over 200 specific application software packages. And since our workstations use HP-UX—our enhanced version of the UNIX* operating system—we'll be able to bring you a lot more in the future.

Of course, when you tie all these individually-optimized workstations together into a network, you can look for an even higher level of performance. Whether you choose our HP-UX networking, Ethernet local area network, or Shared Resource Manager, your engineers can share data and peripherals, communicate much better, and get more done than ever before.

The HP Engineering Productivity Network. You can start with one workstation, or with dozens. But no matter where you start, you're actually making your whole department more productive. One step at a time.

To find out more, just call the nearest HP office listed in the white pages of your telephone book, and ask for a Technical Computer Representative. Or write to Pat Welch, Dept. 41189, Hewlett-Packard, 19447 Pruneridge Avenue, Cupertino, CA 95014. In Europe, write Henk van Lammeren, Hewlett-Packard, Dept. 41189, P.O. Box 529, 1180 AM, Amstelveen, the Netherlands.

We'll be happy to show you how far you can really go with a good idea.

*UNIX is a trademark of Bell Laboratories.

HEWLETT PACKARD

BD-02316

The DNA Helix and How It Is Read

X-ray analysis of crystals of three types of double-helical DNA molecule leads to the realization that base-sequence information can be stored in the local structure of the helix

by Richard E. Dickerson

Two kinds of genetic information are carried in the double helix of DNA, and they are encoded and interpreted in very different ways. The genetic message itself—the information specifying the structure of proteins—is written in the familiar triplet genetic code. That code is linear, and its interpretation is extrinsic. The sequence in which successive triplets of the chemical groups called bases are arrayed along a single strand of the helix encodes the sequence in which successive amino acids are linked to form a protein chain. There seems to be no innate structural relation between a particular base triplet and the amino acid it specifies. Rather, information transfer is mediated indirectly by complex machinery extrinsic to the DNA: messenger RNA, ribosomes, transfer RNA and a battery of attendant enzymes.

DNA encodes not only the genetic message itself but also instructions for the selective expression of the message. What little is known about this kind of information comes from studies of genetic control in bacteria. Blocks of genes on the bacterial chromosome are turned off by the binding of a repressor protein to a region of DNA (the operator) that has a special base sequence the repressor recognizes; the genes become active when various small molecules bind to the repressor and knock it off the operator. Although much remains to be learned, most molecular biologists expect comparable mechanisms to be the basis of genetic control in life forms higher than bacteria. A bacterial repressor recognizes its operator directly, probably by forming hydrogen bonds between nitrogen and oxygen atoms on its amino acids and on the edges of the complementary pairs of bases in the double helix of DNA. The helix fits snugly against a particular conformation of the protein. The information specifying genetic control, then, is three-dimensional, and its interpretation is intrinsic: it depends on the implicit structural properties of the protein and the helix.

Does DNA play only a passive role in the recognition process, with amino acid side chains inserting themselves into the grooves of a static double helix, or does the base sequence itself modify the structure of the helix in a way that contributes to the recognition of specific sites by the control proteins? As recently as four years ago the question would have been pointless because it was totally unanswerable from the type of data then available.

James D. Watson and Francis Crick proposed their double-helix structure for DNA in 1953 on the basis of X-ray-diffraction photographs made by Rosalind Franklin and Maurice Wilkins. The photographs were of fibers of DNA. Further analysis of such diffraction patterns led to the realization that fibers of DNA can exist in two forms: as *B* DNA under conditions of high humidity and as *A* DNA when the humidity is lower. Molecular models were developed for both forms. The amount of structural information that could be extracted from fiber diffraction patterns was inherently limited, however, by the disorder of the DNA strands around and along the fiber axis. The most that could be established was an averaged, overall helix structure. Any local variations in structure, such as might be induced by a particular base sequence, could not be detected.

Now they can be detected and measured precisely. Improved methods for the organic synthesis of short DNA molecules of any desired sequence have made it possible for the first time to make DNA oligomers (short DNA chains) with from four to 24 bases in sufficient quantity and pure enough to be crystallized and studied by conventional single-crystal X-ray-diffraction methods. This article is intended as a window on some of the exciting new results beginning to come from the structural analysis of short double-helical DNA molecules.

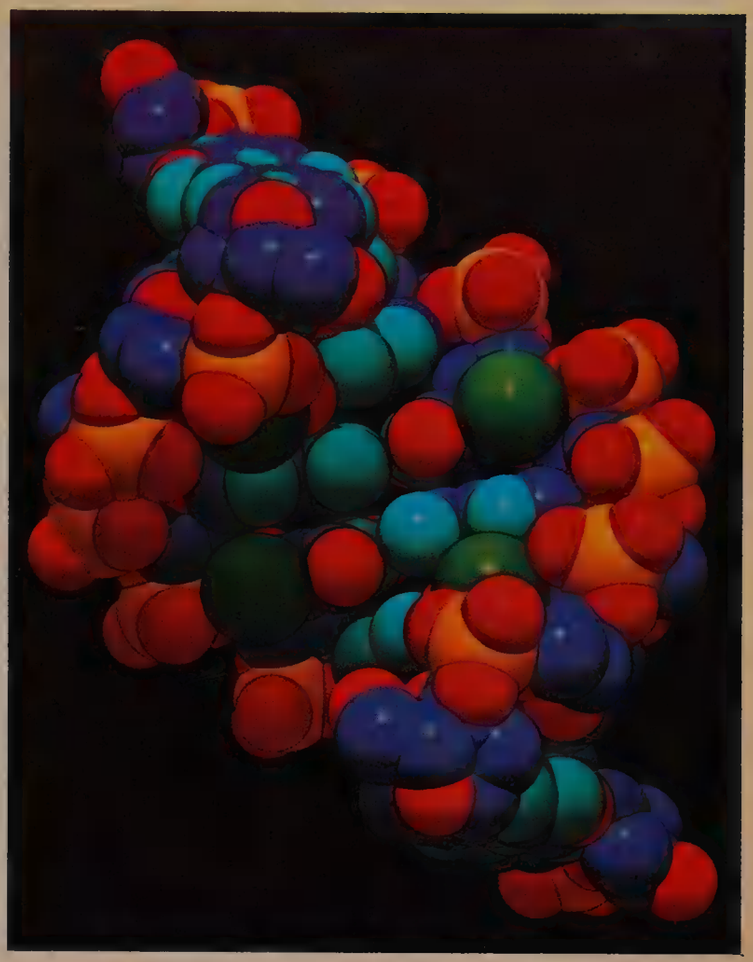
Fiber Diffraction

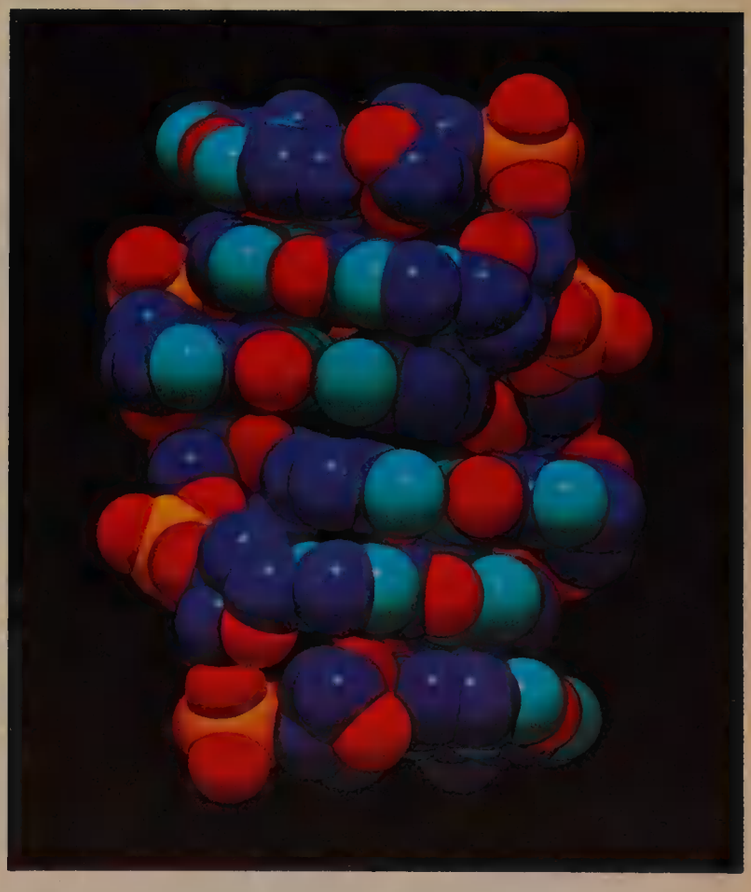
Examination of X-ray patterns from stretched fibers and thin films of natural DNA revealed the basic structure of the two fundamental types of double helix: the *B* configuration, which is stable at a relative humidity of about 92 percent, and the *A* configuration, which most base sequences assume when the humidity falls to about 75 percent. Both have the shape of a flexible DNA ladder wrapped helically around a central axis. The two rails of the ladder are chains of alternating deoxyribose sugar rings and phosphate groups. The rungs are purine-pyrimidine base pairs. There are two kinds of double-ring purine bases, adenine (*A*) and guanine (*G*), and two kinds of single-ring pyrimidines, thymine (*T*) and cytosine (*C*). The base pairs are held together by hydrogen bonds. Adenine normally is paired with thymine by two hydrogen bonds and guanine is paired with cytosine by three bonds.

The two points of attachment of a base pair to its sugar rings do not lie directly opposite each other across the pair [*see bottom illustration on page 97*], and this is important for the geometry of a DNA double helix. The edge of the base pairs along which the angle between attachments is less than 180 degrees is called the minor-groove edge, and the one forming an angle larger than 180 degrees is the major-groove edge. When base pairs are stacked into a helix, the phosphate backbones build the two walls of a major and a minor groove, which wind around the helix with these base edges as the floor of the grooves.

The floor of the major groove is paved with nitrogen and oxygen atoms that can make hydrogen bonds with the side chains of a protein's amino acids and therefore can have a significant role in







THREE TYPES OF DOUBLE HELIX are displayed in these space-filling models, which are computer images generated by Nelson Max of the Lawrence Livermore National Laboratory. They show three of the short double-helical molecules whose structures have been solved by single-crystal X-ray-diffraction analysis. Carbon atoms are dark blue, nitrogen atoms light blue, oxygen atoms red and phosphorus atoms yellow. Bromine atoms on some of the bases are green. *B* DNA (*left*) is represented by a 12-base molecule whose structure was solved by the author and his colleagues. The view is directly into the major groove of the helix. The phosphate backbone gives a strong sense of the right-handed twist. The eight-base molecule of *A* DNA at the top right is one solved by Olga Kennard, Zippora Shakked and M. A. Viswamitra. It too is right-handed. The view, into the very deep major groove, shows how the base pairs are tilted with respect to the vertical axis of the helix. The molecule of *Z* DNA at the bottom right is one whose structure was solved by Andrew H.-J. Wang, Alexander Rich and their colleagues. As the phosphate groups at the upper right and the lower left of the image indicate, the *Z*-DNA helix is left-handed. The backbone zigzags, and the base pairs are stacked by twos rather than singly, giving the helix an alternating structure.

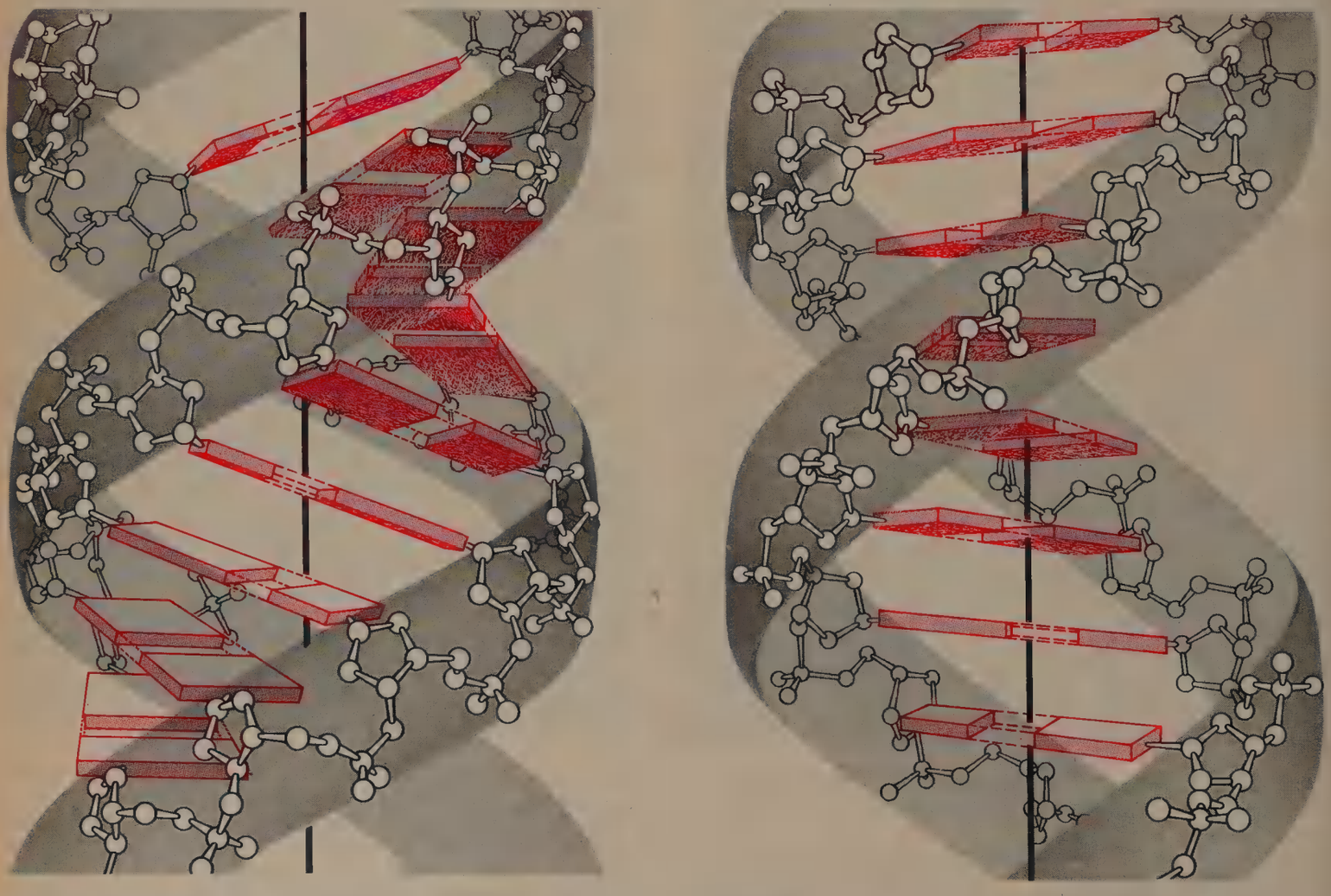
intrinsic coding. The pattern of these hydrogen-bonding groups is different for the two kinds of base pair. Reading from purine to pyrimidine, an *A–T* pair offers a nitrogen atom (a hydrogen acceptor), an NH_2 group (a donor) and an oxygen atom (another acceptor). In contrast, the *G–C* pair offers the same groups in a different order: first a nitrogen (an acceptor), then an oxygen (an acceptor) and finally an NH_2 (a donor). Since each base pair (*A–T* and *G–C*) can also be turned around (to become *T–A* and *C–G*), four different patterns can be exhibited to a repressor or another control protein at each step of the helix. The bottom of the major groove therefore carries a message, the base sequence of the DNA, in a form that can be read by other large molecules.

The minor groove is less informative: the hydrogen-bonding pattern in *A–T* base pairs is simply acceptor-acceptor no matter which way the base pair is flipped, and a *G–C* pair differs only by the intrusion of an NH_2 donor between the acceptors. This makes the minor groove a poorer candidate for information readout. As will emerge below, the minor groove has another important function in *B* DNA.

A DNA and *B* DNA differ mainly in the positioning of the base pairs around the helix axis and in base inclination: the tilt of the bases with respect to the axis. In *B* DNA the inclination is close to zero (the base pairs are stacked nearly perpendicular to the helix axis) and the axis runs through the center of each base pair. Although the minor groove is narrower than the major groove because of the base pairs' asymmetric attachment to the sugar rings, the grooves are of similar depth: roughly the same distance must be traversed in from the surface of the helix cylinder before the edge of a base pair is encountered. In *A* DNA, on the other hand, the base pairs are tilted between 13 and 19 degrees from the perpendicular. Moreover, they are shifted toward the outside of the helix, whose axis lies in the major groove, bypassing the bases. Hence the minor groove is shallow, scarcely more than a helical depression spiraling around the outside of the cylinder; the major groove is cavernously deep, extending from the surface all the way past the central axis and part of the way out toward the opposite side.

B DNA has an average of 10 base pairs per turn of the helix, with a spacing of 3.4 angstrom units along the helix axis from one base pair to the next. The *A* helix has closer to 11 base pairs per turn, and because of the way the tilted base pairs are stacked, the distance along the axis is only 2.9 angstroms per base pair. These distances are averages, derived from fiber data; single-crystal analysis eventually revealed large local deviations from the average values. (The original Watson-Crick model had the 10 base pairs per turn and the perpendicular stacking of bases that is characteristic of *B* DNA, but the bases were pulled away from the center to leave a central hole more characteristic of the *A* form. The positioning of sugar rings in the original model also was more like that of the *A* form.)

This was pretty much the state of knowledge about double-helical DNA structure for 25 years after Watson and Crick. Many questions remained unanswered and unanswerable. Why is the *B* helix the high-humidity form? Which form is prevalent in chromosomal DNA, and can one form be converted into another in a living organism? Do recognition proteins such as repressors actually read information from the bottom of the grooves when they bind to specific DNA sequences?



ANALYSES OF DNA FIBERS by Struther Arnott and others, done before the single-crystal studies reported in this article, were the basis of these schematic drawings of *A* DNA (*left*) and *B* DNA (*right*) made by Irving Geis. The sugar-phosphate backbones of the double helix are represented as ribbons and the runglike base pairs connecting them as planks. In *A* DNA the base pairs are tilted and are pulled away from the axis of the double helix. In *B* DNA, on the other hand, the base pairs sit astride the helix axis and are perpendicular to it.

In the late 1970's, when improvements in DNA synthesis made it possible for the first time to study single crystals of short molecules of any selected sequence, both Alexander Rich's group at the Massachusetts Institute of Technology and our own group, which was then at the California Institute of Technology, independently undertook to synthesize *C–G* copolymers: short molecules composed solely of alternating cytosines and guanines. There was a reason for the choice of this sequence. Experiments in solution suggested that the *C–G* copolymer undergoes some kind of structural transition under high-salt or high-alcohol conditions, and so it was the most likely candidate for studying transitions between helix types.

Single-Crystal Studies

Andrew H.-J. Wang of M.I.T. solved the structure of the hexamer *CGCGCG* in mid-1979 and Horace R. Drew in our laboratory determined the structure of the tetramer *CGCG* a few months later. To everyone's surprise the two molecules proved to be neither *A* DNA nor *B* DNA. They were not even right-handed helixes. Instead they were left-handed helixes of an entirely new type, with a peculiar zigzag backbone that led them to be named the *Z* helix. Fiber studies had prepared us to ask a well-defined question: Are these molecules in the *A* form or the *B* form? Nature gave us the chastening and almost rude reply: No.

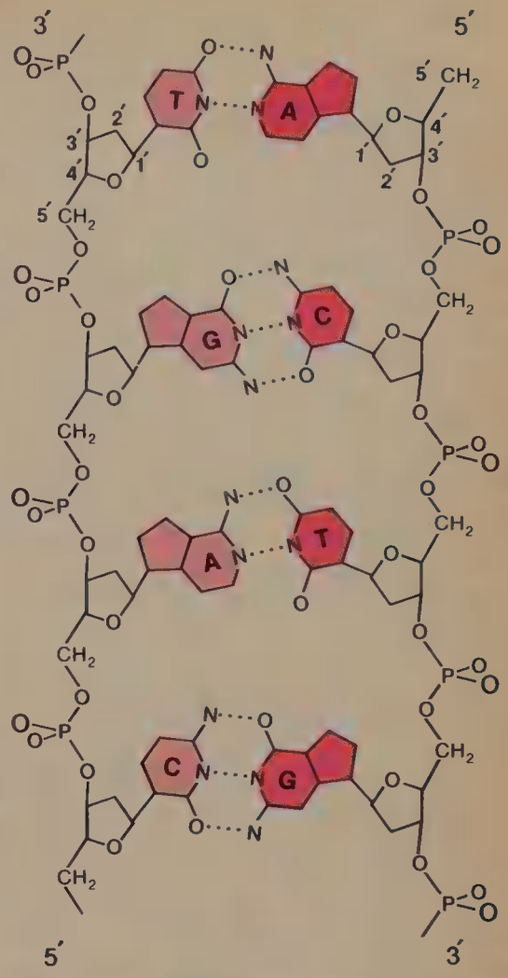
Although the first two double-helical DNA molecules solved by single-crystal X-ray methods broke with all expectations, the next two provided one example each of *B* and *A* DNA. Drew synthesized the dodecamer *CGCGAATTCGCG*, a polymer we chose for two reasons. It coupled *Z*-compatible *CGCG* ends with a *Z*-incompatible *AATT* center, providing a test of the *Z*-forming power of *CGCG* in alien surroundings. And it included the sequence *GAATTC*, which was of interest because it is the recognition and cleavage site for *Eco RI*, one of the restriction endonucleases: enzymes that cut DNA molecules at specific sites. (The reader may note that the oligomer *CGCGAATTCGCG*, like others synthesized for these studies, is self-complementary: the sequence read from left to right is complementary to the sequence read from right to left, so that any two such chains can combine to form a double helix. The result is a major saving of time and energy, since only one strand of a double helix need be synthesized.) The structure of the *CGCGAATTCGCG* molecule turned out to have nothing to do with *Z* DNA but rather to provide a classical example of a *B* helix. Benjamin N. Conner in our laboratory also synthesized the tetramer *CCGG*, solved its structure and showed it to be a fragment of *A* DNA.

Other double-helical DNA molecules followed in quick succession from several laboratories. The *A* helix is represented by *CCGG*, by *GGCCGGCC* from the M.I.T. group and by *GGTATACC* from a group that includes Olga Kennard at the University of Cambridge and Zippora Shakked at the Weizmann Institute of Science in Israel. Wang has also solved the structure of an *A* DNA molecule that is a mixed RNA-DNA hybrid double helix: *(GCG)TATACGC*. (The letters enclosed in parentheses represent RNA bases, so that the outer three base pairs, at each end of the double helix, are RNA-DNA hybrids and the central four base pairs are pure DNA.) *B* DNA is represented so far only by *CGCGAATTCGCG* and its derivatives. *Z* DNA is found in the above-mentioned *CGCG* and *CGCGCG* and in a variant of the latter in which methyl groups (CH_3) are added to cytosines, a modification favoring the *Z* structure. These molecules now can be analyzed to see how well they match predictions for *A* and *B* DNA from fiber diffraction, how much variation there is from average helix properties and to what extent the variation can be attributed to the base sequence of the DNA itself.

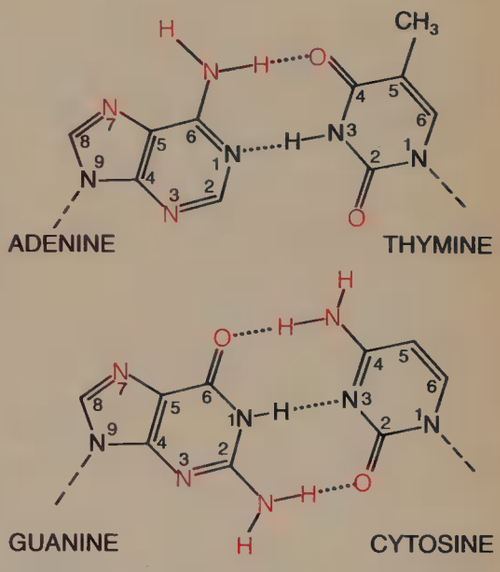
Structural Features

Average values for structural features of each kind of helix have been determined from single-crystal analyses [*see table on next page*]. *A* and *B* DNA are right-handed, with average helical-twist angles from one base pair to the next of 33.1 and 35.9 degrees respectively, corresponding to 10.9 and 10.0 base pairs per 360-degree turn, in close agreement with predictions based on fiber studies. The rise per base pair along the helix axis also has no surprises. What is surprising is the extent of variation in helical-twist angles around these average values: a standard deviation of ± 6 degrees for *A* DNA and of ± 4 degrees for the *B* form. Individual twist angles are as small as 16 degrees and as large as 44 in *A* DNA, and they range from 28 to 42 degrees in *B* DNA. This variation in local helix rotation can be predicted directly from the base sequence by a method I shall describe below. (Is this variable one element by which repressor and other control molecules recognize particular base sequences? The idea is an attractive one.) The bases are inclined away from perpendicularity to the helix axis roughly as was expected from fiber analyses: 13 degrees in *A* DNA and about two degrees in *B* DNA.

A greater departure from fiber-based expectations is found in the "propeller twist" of individual base pairs. Propeller twist is a rotation of the two bases of a pair in opposite directions about their long axis; the sense of the twist is said to be positive when the nearer base (look-



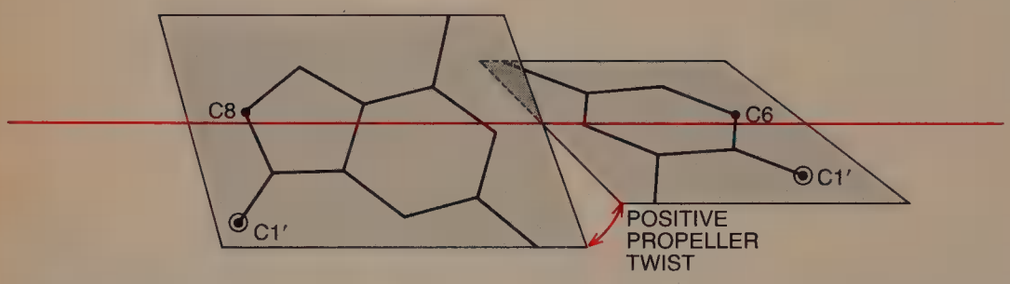
HELIX IS UNROLLED to diagram DNA's sugar-phosphate backbones and base-pair rungs. The backbones run in opposite directions, with 5′ and 3′ ends named for the orientation of the 5′ and 3′ carbon atoms of the sugar rings. Each base pair has one purine base, adenine (*A*) or guanine (*G*), and one pyrimidine base, thymine (*T*) or cytosine (*C*), connected by hydrogen bonds (*dotted lines*).



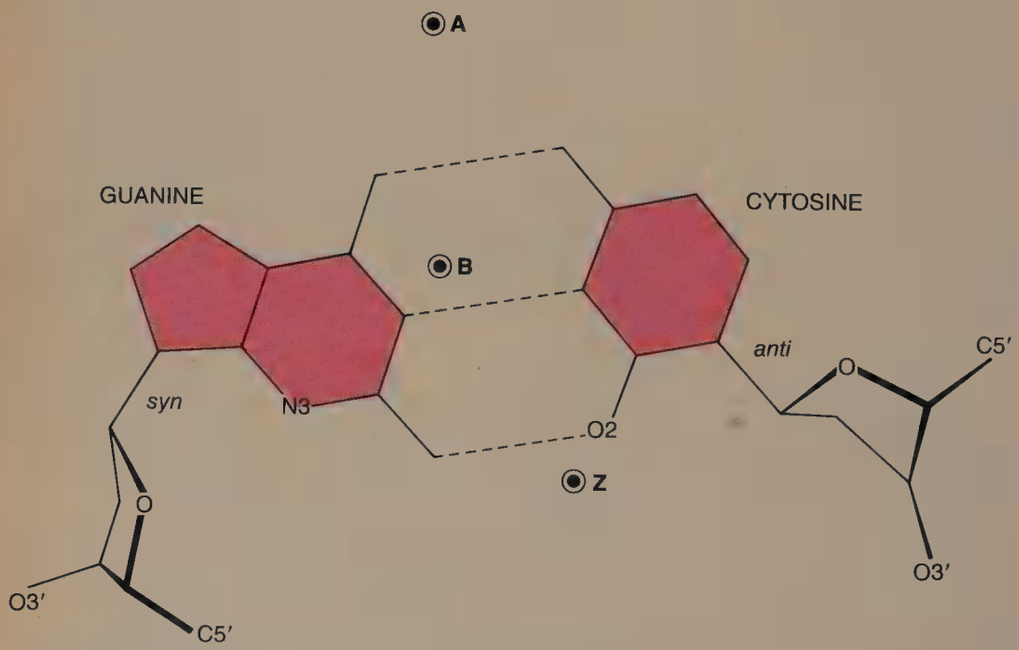
BASE PAIRS are shown in detail, with broken lines indicating bonds to sugar rings of the backbone. Each base pair has a major-groove (upper) edge and a minor-groove (lower) edge. Amine groups (NH_2) are potential hydrogen donors in forming hydrogen bonds with repressors or other regulatory molecules; the nitrogen and oxygen atoms indicated in color are potential hydrogen acceptors. A base pair's characteristic pattern of donors and acceptors is probably "read" by control proteins.

	A DNA	B DNA	Z DNA
HANDEDNESS	RIGHT	RIGHT	LEFT
HELICAL TWIST (DEGREES)			
MEAN AND STANDARD DEVIATION	33.1 ± 5.9	35.9 ± 4.3	*G–C:* −51.3 ± 1.6 *C–G:* −8.5 ± 1.1
OBSERVED RANGE	16.1 to 44.1	27.7 to 42.0	
BASE PAIRS PER TURN	10.9	10.0	12.0
HELIX RISE PER BASE PAIR (ANGSTROM UNITS)	2.92 ± .39	3.36 ± .42	*G–C:* 3.52 ± .22 *C–G:* 4.13 ± .18
BASE INCLINATION (DEGREES)	13.0 ± 1.9	−2.0 ± 4.6	8.8 ± .7
PROPELLER TWIST (DEGREES)	15.4 ± 6.2	11.7 ± 4.8	4.4 ± 2.8
BASE ROLL (DEGREES)	5.9 ± 4.7	−1.0 ± 5.5	3.4 ± 2.1

MEAN HELIX PARAMETERS and standard deviations are listed for three kinds of DNA. The data are based on single-crystal X-ray analysis of molecules mentioned in this article.



PROPELLER TWIST is defined by a schematic drawing of a purine-pyrimidine base pair. A clockwise rotation of the nearer base as one sights down the long axis (*color*) of the base pair in either direction is considered positive propeller twist. Note that the C1′ atoms, at which bases are attached to sugar-phosphate backbones, are shifted up and down by the propeller twist.



TWO CONFORMATIONS of the C–N bond connecting each base to its sugar ring are shown. The *anti* conformation (*right*) appears in all *A* and *B* DNA and at the cytosines in *Z* DNA; the *syn* conformation (*left*) appears at the guanines in *Z* DNA. The alternation of a purine and a pyrimidine (*G, C*) along a strand of *Z* DNA makes possible the *syn-anti* alternation that causes the *Z*-helix backbone to zigzag. The three points labeled *A, B* and *Z* indicate for each type of helix the location of the helix axis with respect to a base pair. Axis is in the major groove in *A* DNA, passes through the base pair in *B* DNA and is in the minor groove in *Z* DNA.

ing along that axis) is rotated clockwise [*see middle illustration at left*]. Little attention had been paid to propeller twist in the fiber studies. A set of coordinates for *A* DNA published in 1972 implied negative propeller twist, and an unpublished revision in 1981 had the proper twist direction but a magnitude of only eight degrees.

Single-crystal studies show that the twist always is positive, with mean values of 15 degrees for *A* DNA and 12 degrees for *B;* individual values range from as little as three degrees to as much as 25. In a right-handed helix positive propeller twist improves the stacking of bases along each individual backbone strand, making each base overlap more with its neighbors up and down the chain. On the other hand, positive propeller twist also leads to uncomfortably close contact between purines (*G* and *A*) on opposite backbone strands at adjacent base pairs. This, as C. R. Calladine of Cambridge has suggested, is chiefly responsible for several important sequence-dependent variations in helix structure, including propeller twist itself, local helical twist and what is called the base-roll angle.

Base roll measures the orientation of the base pair as a whole (the best mean plane through the purine and the pyrimidine) about its long axis. If two successive base pairs along the helix are rolled in opposite directions, they open up an angle between them toward either the major or the minor groove. The roll angle from one base pair to the next is defined as positive if the angle between them opens toward the minor groove and as negative if it opens toward the major groove. There is no way to measure local base-roll angles from fiber-diffraction data, but single-crystal studies show that base roll is one of the important helix parameters. Since in *B* DNA the base pairs are nearly perpendicular to the helix axis, there need be no systematic mean roll angle from one base pair to the next. In the *A* form, however, the DNA ladder is wrapped around the helix axis in a way that requires an accordionlike opening of base pairs toward the minor groove, and the mean roll angle in *A* DNA is six degrees. Even more significant is the variation, in both the *B* and the *A* helix, of about ±5 degrees from these mean values.

Expansion to Infinite Helixes

The longest DNA double helix that has yet been examined by single-crystal X-ray methods is the dodecamer *CGCGAATTCGCG.* An impression of what longer stretches of helix would look like can be gained by extending three typical molecules [*see illustration on page 95*] to produce long helixes [*see illustrations on pages 100 through 105*].



CIMARRON '84
THIS ONE'S GOT THE TOUCH.

The Cadillac touch. It's everywhere. You can feel it when you put Cadillac's road-hugging Touring Suspension to the test. You can see it ...inside and out. From leather-faced front buckets with lumbar support to aluminum alloy wheels. It's attention to detail, too. Like push-button air. And tungsten halogen fog lamps. You've got to drive this car... and experience the Cadillac touch. Cimarron '84.

BEST OF ALL...IT'S A CADILLAC.

Let's Get It Together...Buckle Up.



This is done by programming a computer to rotate an image of a given helical molecule along its axis until atoms near the beginning of the shifted image coincide with equivalent atoms near the end of the original image, and continuing the process for as long as desired. In this way, for example, the central six base pairs of *A*-helical *GGTATACC* have been repeated four times to produce a 24-base-pair helix with the sequence *GTATACGTATACGTATACGTATAC*. For the first time it has been possible to generate a long double helix from molecules obtained by single-crystal X-ray analysis, without averaging out or idealizing the local, sequence-induced fluctuations in helix structure.

As was expected from fiber studies, the *A* helix is short and fat, with a deep major groove and a broad and shallow minor groove. The *B* helix is slimmer (and taller for the same number of base pairs), with a wide major groove and a narrow minor groove of comparable depth. In contrast, the new left-handed *Z* helix is quite thin and elongated, with a deep, narrow minor groove and a major "groove" that is pushed to the surface so that it no longer is a true groove at all. Because the sugar-phosphate helix backbone in *Z* DNA follows a zigzag path, the repeating unit of the helix is not a single base pair, as it is in *A* and *B* DNA, but rather two successive base pairs: *G–C* followed by *C–G*. This arises because of a difference in the way cytosines and guanines are attached to their sugar rings in *Z* DNA.

At every cytosine the sugar is rotated about its bond to the base so that the puckered sugar ring swings away from the minor groove [*see bottom illustration on page 98*]. This conformation, designated *anti,* also is present in the case of all four bases, and therefore at every step, in *A* and *B* DNA. Each guanine in *Z* DNA, however, has its sugar ring rotated 180 degrees, so that it bends inward toward the minor groove. This *syn* conformation is stereochemically possible only when the sugar is attached to the smaller, five-member ring of a purine rather than to the six-member ring of a pyrimidine; the distance between a *syn* sugar ring and the oxygen atom at position O2 of a cytosine or thymine ring would be too short to be stereochemically allowable. It is the alterna-

A-DNA HELIX is seen in this perspective drawing based on the computer-generated stereoscopic pair on the opposite page. The stereo pair was generated by extending the central six bases of the octamer *GGTATACC*, the structure of which was determined by Kennard, Shakked and Viswamitra. The base pairs are not idealized; they are as they are observed in the crystal-structure analysis. Note how phosphate groups on opposite chains face each other across major groove.

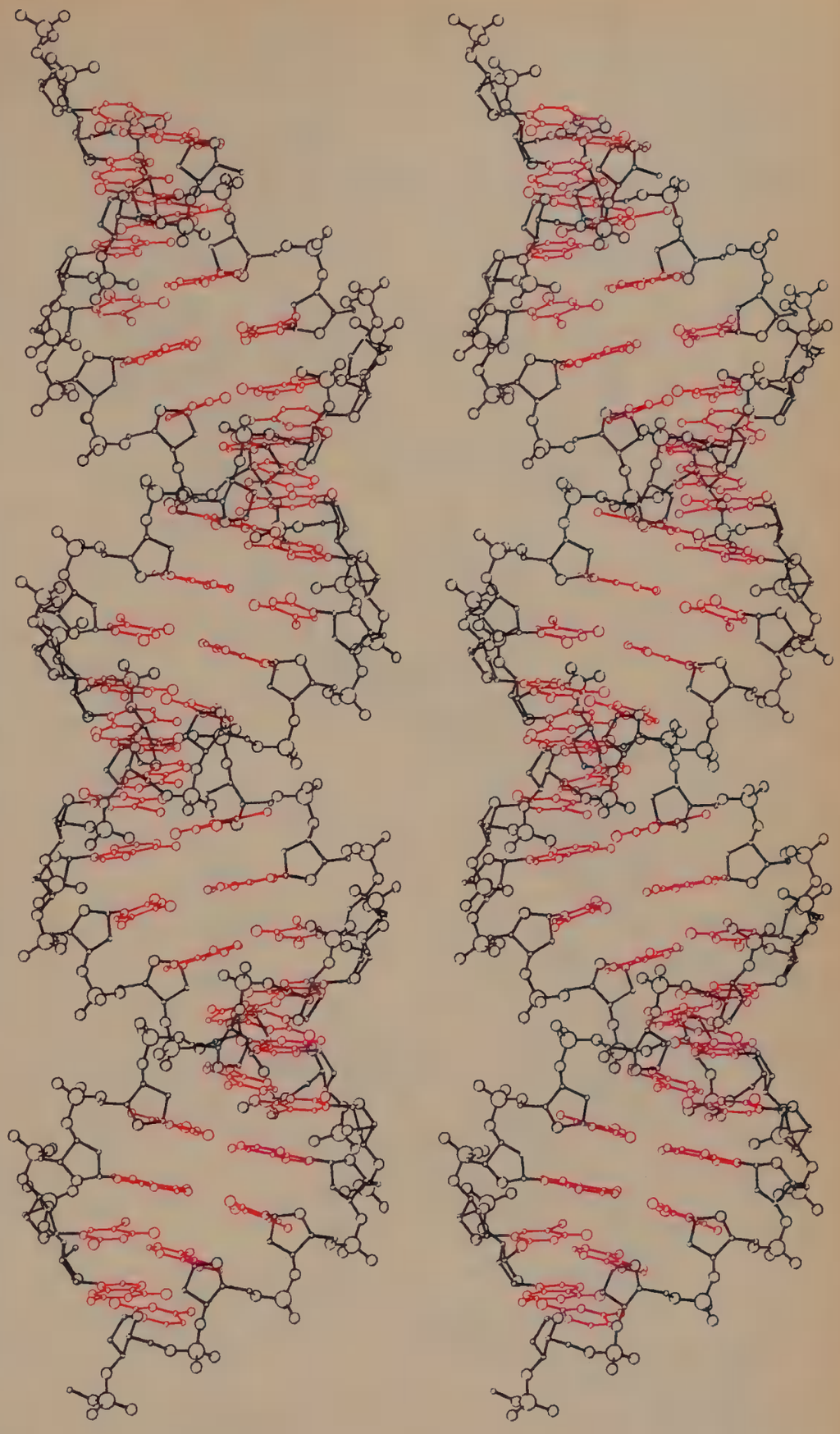
tion of a purine (*G*) and a pyrimidine (*C*) along each chain in *Z* DNA that allows a *syn-anti* alternation of bonds between sugars and bases and produces the zigzag backbone chain. The path of the chain from one phosphorus atom past a guanine to the next phosphorus lies nearly parallel to the helix axis; the path from phosphorus to phosphorus past a cytosine, on the other hand, is tangential to the cylindrical envelope of the helix and perpendicular to its axis. As a result the true repeat unit along the helix is two successive base pairs; although the *Z* helix has 12 base pairs per turn, it formally is a sixfold left-handed helix with six sets of two base pairs per turn.

RNA

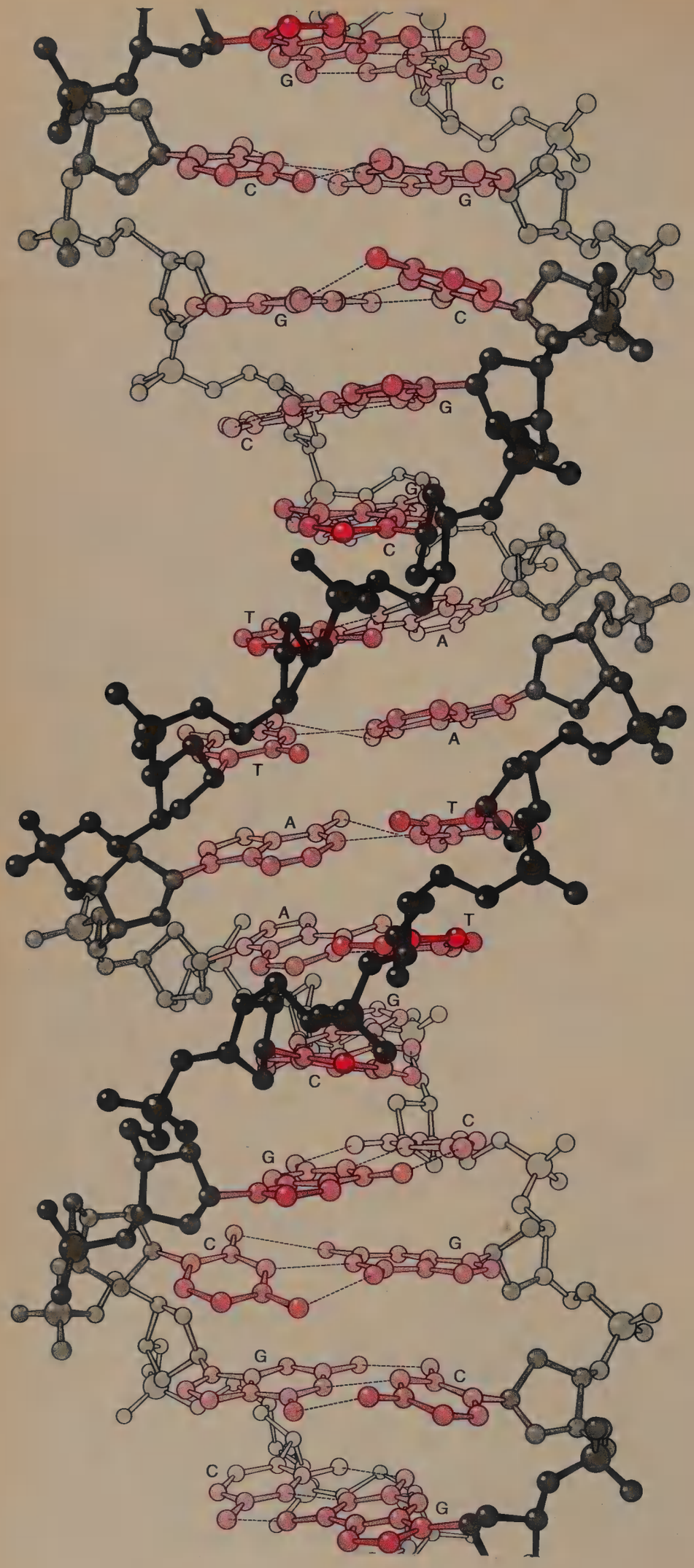
So far I have dealt only with DNA (deoxyribonucleic acid), the archival storage material for genetic information in the nucleus of the cell. The expression of that information depends in large part on the closely related nucleic acid RNA (ribonucleic acid). The base sequence on one strand of the DNA helix is transcribed into a single strand of messenger RNA; transfer-RNA molecules read the sequence and bring the designated amino acids to the ribosomes for assembly into a protein chain; the ribosomes themselves are composed partly of RNA. There are just two differences between DNA and RNA. DNA's thymine is replaced in RNA by uracil, which lacks thymine's methyl side chain, and instead of the deoxyribose sugar rings of DNA's backbone the RNA backbone has ribose rings. In deoxyribose the hydroxyl (OH) group attached to one carbon atom (C2′) of the ribose ring is replaced by a hydrogen.

An important consequence of the extra hydroxyl in RNA is that RNA apparently cannot coil into a *B* double helix. In a hypothetical *B* RNA the OH group would sit in the middle of a cage of atoms of the following phosphate group, sugar ring and base. The distance between the extra oxygen atom and several other atoms would be unacceptably short, making the structure stereochemically unfavorable. In *A* DNA, on the other hand, the 2′ OH group projects out from the helix surface, away from any nearby atoms.

The self-complementary double-helical loops and coils in the transfer-RNA molecule must therefore be variations of an *A* helix. If, as has been suggested, short self-complementary regions of single-strand messenger RNA can form hydrogen-bonded hairpin loops, then these too must be *A* RNA rather than *B*. (In the only example known of RNA participation in a *B* helix, a synthetic hybrid with adenines on the RNA strand and thymines on the DNA strand, the intrusive OH is accommodated by a hy-



STEREO PAIR was generated by manipulating the three-dimensional set of coordinates representing an image of the double-helical *A*-DNA hexamer *GTATAC*, which was derived by deleting the base pairs at the ends of the octamer (because they lacked phosphate groups). The image was rotated and translated along the helix axis until the bottom pair of phosphates on the image coincided with the top pair on an unshifted image; the process was repeated several times to yield what is in effect an infinite *A* helix. The method makes it possible to generate a long stretch of a specific DNA without sacrificing sequence-determined local variations in helix structure. A stereo pair is most easily seen in perspective through a magnifying-lens stereo viewer. With practice a three-dimensional image can be seen without a viewer if one develops the ability to decouple two reflexes that are normally associated: to deconverge one's eyes as if looking at a distant object while focusing the eyes for near vision. Note the depth of the major groove. The helix resembles a double-strand ribbon wound around the outside of a cylinder.



drogen bond to the O4′ atom of the following ribose.) No one has seriously suggested the possibility of a *Z* RNA, but a close examination of the structure indicates that an OH added to a cytosine sugar probably would lie too close to the projecting O2 atom of the cytosine ring.

Stabilization by Hydration

Single-crystal X-ray structure analysis has defined three basic types of DNA double helix rather than two, and one type of RNA double helix. For DNA the *B* form is the stable one under conditions of high humidity, and this makes it seem that the *B* helix is most likely to be present in the nucleus of the cell. If a transient hybrid DNA-RNA helix is formed in the course of the transcription of DNA into messenger RNA, it probably adopts the *A* structure. Under special conditions of stress, right-handed DNA with the purine-pyrimidine alternating sequence can be flipped over into the left-handed *Z* state. For general sequences and under aqueous conditions, however, *B* DNA seems to be the norm. How does water stabilize the *B* helix?

After a DNA molecule has been refined by X-ray analysis one can make a cautious search for solvent (water) molecules around it that happen to be highly ordered, that is, so strongly bound to the DNA that they occupy the same places around every molecule in the crystal. (Less tightly bound water molecules that are at different sites from one molecule to the next in the crystal are blurred in the diffraction image and are barely distinguishable from the background noise.) The water structure has been examined in molecules of all three forms of the DNA double helix, but detailed results have been published only for *B* DNA. Water molecules are observed in *B* DNA in the vicinity of nearly every atom that could make hydrogen bonds with them: free phosphate oxygens, nitrogen and oxygen atoms on the edges of base pairs and, to a lesser extent, the phosphate oxygens that are part of the backbone chain. In effect the helix is coated with a layer of water one molecule deep.

A more extended hydration structure is observed within the narrow minor groove of *B* DNA [*see bottom illustration on page 109*]. The combination of the propeller twist of individual base pairs and the helical twist from one base pair to the next brings an oxygen or nitrogen of one chain into close proximity to an

B-DNA HELIX depicted in this drawing was generated from repetition of the central 10 base pairs of the dodecamer *CGCGAATTCGCG*. The drawing shows the large propeller twist and how the twist improves the stacking of the bases along each backbone chain.

oxygen or nitrogen of the other chain on an adjacent base pair. The two hydrogen-bond acceptors are bridged by a water molecule. These first-layer molecules are bridged in turn by a second layer of water molecules. Together the two sets of molecules build a zigzag spine of hydration running down the minor groove as it spirals around the *B* helix. The spine of hydration is well developed in regions of *A–T* base pairs but is broken up in *G–C* regions, probably because of the intrusion into the minor groove of NH_2 groups on guanines. In synthetic DNA fibers any base pairs in which guanine is replaced by the modified base inosine, which lacks guanine's NH_2 group, behave like *T–A* pairs rather than *C–G* pairs, in that they help to stabilize the *B* structure.

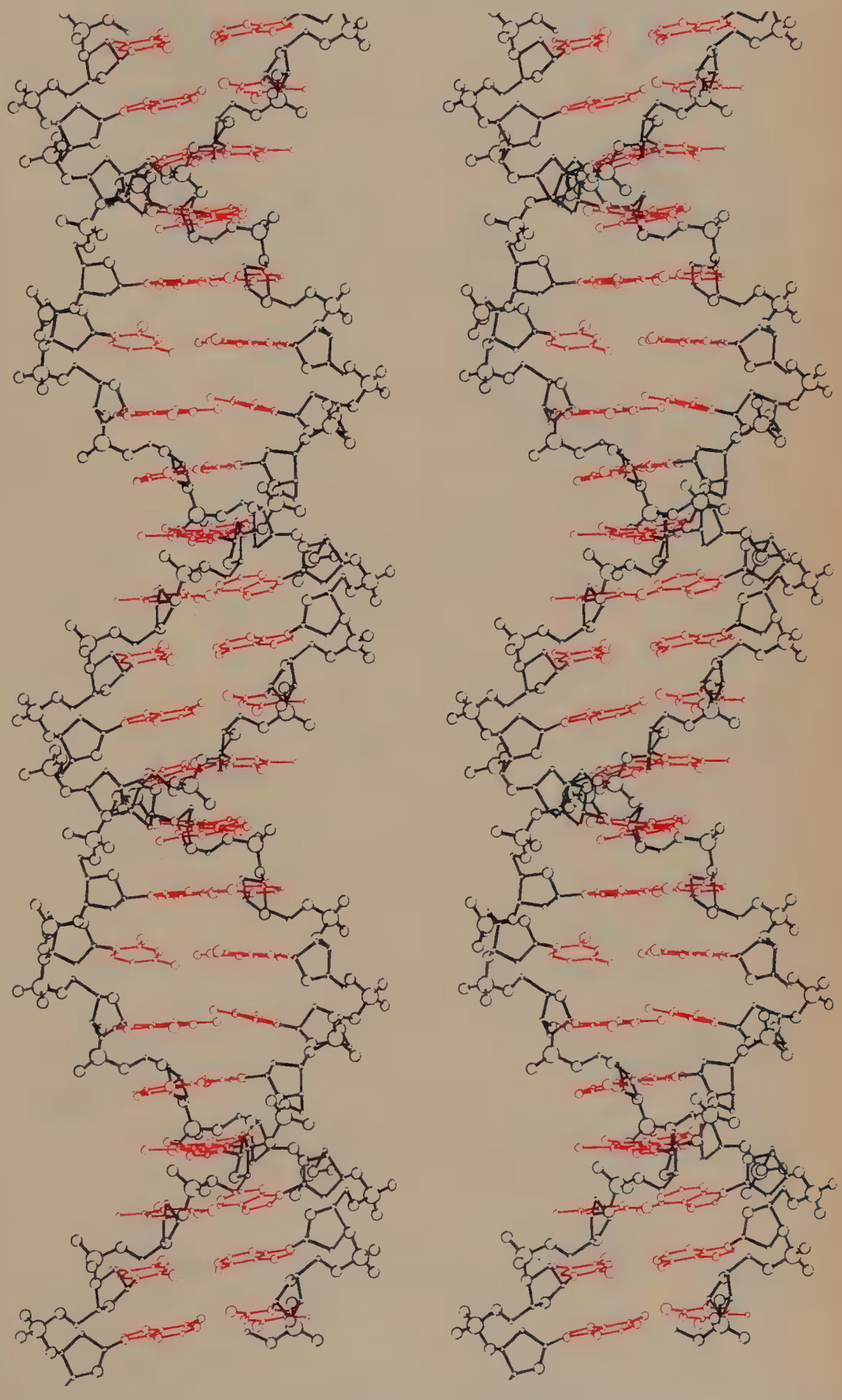
We think that this minor-groove spine of hydration is a prime stabilizing influence on *B* DNA, and that it must be physically disrupted before the minor groove can open up and the helix can slip over into the *A* form. One way to disrupt the stabilizing spine in the laboratory is to remove water molecules by dehydrating the fibers. Another way is to introduce more NH_2 groups into the minor groove; fiber-diffraction studies by Struther Arnott of Purdue University have shown that increasing the *G–C* content in a fiber makes it easier to bring about the transition from the *B* to the *A* form by drying.

No comparable systematic water structure has been found around either the *A* or the *Z* helix. Although water molecules are distributed liberally around atoms on the DNA that could take part in hydrogen bonds, there is nothing that is comparable to the minor-groove spine of hydration in *B* DNA and could be expected to have structural integrity; indeed, the broad minor groove of *A* DNA seems to be even less hydrated than other parts of the *A* helix. It is hard at present to be sure about *A* DNA hydration, however, since in all three *A* DNA crystals—*CCGG*/*CCGG* (two tetramers stacked to make what is in effect an octamer), *GGCCGGCC* and *GGTATACC*—the end base pairs of one octamer pack against the minor grooves of neighbors, blocking some of the normal hydration that would be expected if the molecules were isolated in solution. Even so, one now has a clear molecular picture of the reasons for the *B* form's prevalence under most general aqueous conditions and for the transition to the *A* form as a result of dehydration.

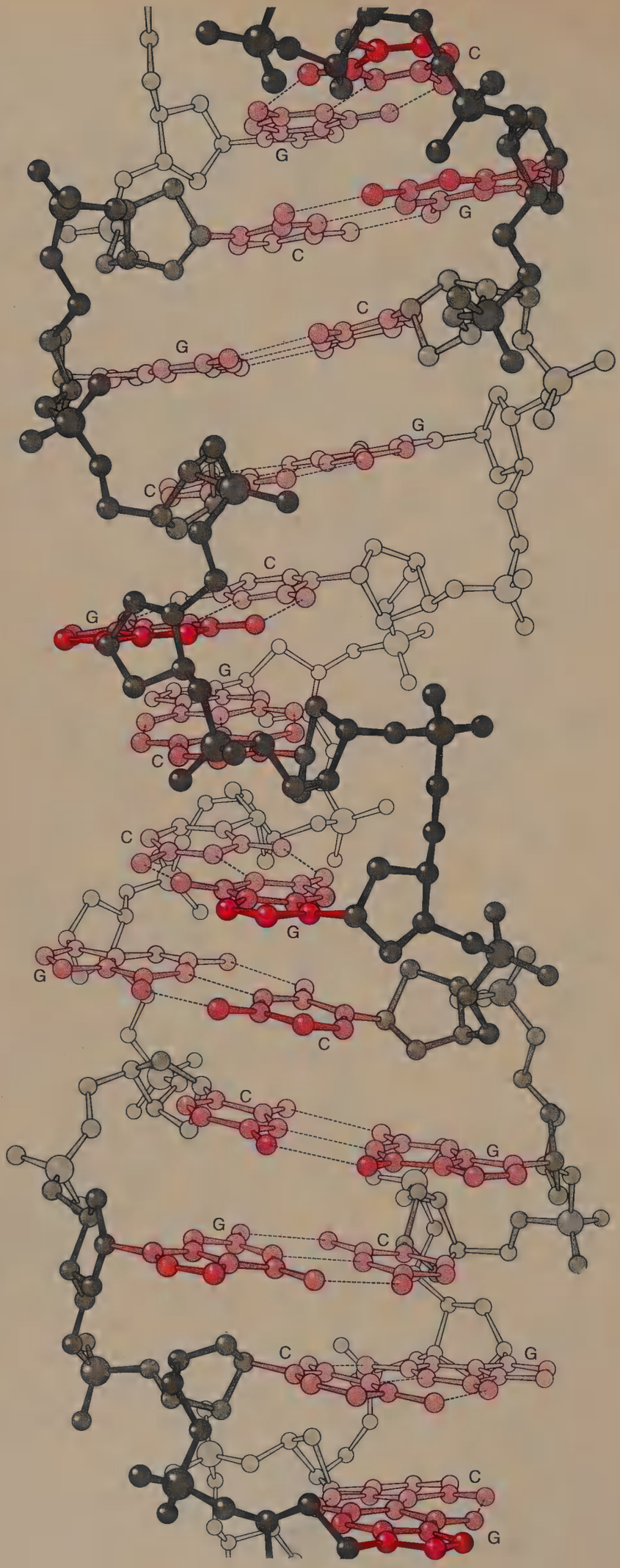
Genetic Control

Let me now turn to what may be the most significant biological implications of single-crystal X-ray studies of DNA: the influence of base sequence on local helix structure and the role this may play in controlling the expression of genetic information. If the *Z* helix is indeed restricted to only a small number of base sequences with alternating purines and pyrimidines, it is unlikely to have much significance in regions of the gene that encode proteins. As Rich and others have suggested, however, the ability of particular regulatory regions of DNA to reverse the sense of their helical twist could be critically important in controlling the accessibility and readout of genetic information, for example by changing the degree of supercoiling of closed circular DNA.

Such a mechanism would be one example of the influence of base sequence not on the content of the genetic message but rather on its expression. The selective expression or repression of genetic information in particular cells at



STEREO PAIR was prepared by deleting the outermost bases of the dodecamer and reiterating the sequence *GCGAATTCGC*. Here one can see that in a *B* helix the major and minor grooves are about the same depth, although the minor groove is narrower. It is particularly narrow in the *AATT* region at the front and is wider at the back in regions rich in *C–G* base pairs.



particular times, however it may be accomplished, is the key to development in higher, multicelled life forms. Moreover, a breakdown in the regulation of genetic information can be one trigger for cancer, and it may be involved in the aging process as well. The current level of understanding of the control of genetic information in higher life forms is approximately at the level of understanding of the coding of that information in 1953, when Watson and Crick first proposed their double helix.

The right-hand-to-left-hand transition between *B* and *Z* DNA represents a gross sequence-dependent conformational change. A different kind of structural influence on the readout of genetic information is being revealed by detailed study of local helix variation in *A* and *B* DNA. As I have indicated, the average values of helix parameters are generally about what had been expected from fiber studies; what was not expected are the large local deviations from these averages.

The mean helical-twist angle for *A* DNA is 33.1 degrees, corresponding almost exactly to 11 base pairs per turn, but who would have predicted that individual values could vary from as little as 16 degrees to as much as 44? The mean base-plane roll angles of six degrees for *A* DNA and nearly zero for *B* DNA might have been expected from the way base pairs are wrapped around the helix axis, but no one expected to find standard deviations from these means of about ± 5 degrees in each case. Such local fluctuations at first caused the crystallographers to reexamine their analyses carefully to ensure that the observed variations were not simply artifacts of the refinement process. With this concern satisfied (and in fact with comparable variations observed in different structures refined in three different ways), the obvious question was whether or not the fluctuations have some predictable basis.

From Sequence to Structure

Several attempts were made to find an explanation for structure variation in terms of the sequence of bases along the DNA. The most successful by far was set in motion by Calladine, who applied principles of elastic-beam mechanics to the double helix. According to Calladine's analysis, local variations in helical twist and base-plane roll arise because of steric hindrance between the

Z-DNA HELIX in this drawing is a left-handed helix of alternating guanines and cytosines, generated from the central four base pairs of *CGCGCG* studied by Wang, Rich and their colleagues. Propeller twist is smaller than it is in the two right-handed helixes. Phosphate groups on different chains now face each other across the deep minor groove.

large purine bases, a consequence of the propeller twist in individual base pairs. (Remember that propeller twist itself improves the stacking of bases along each strand of the double helix.)

Because a double-ring purine (*G* or *A*) is larger than a single-ring pyrimidine (*C* or *T*), it extends past the helix axis in any one base pair. If two purines are on opposite backbone strands at adjacent base pairs, they overlap slightly when they are viewed in projection down the helix axis. If each base pair now is given a positive propeller twist, the edges of the purines are brought into unacceptably close contact and a stereochemical clash results. For a purine-to-pyrimidine sequence progressing in the normal 5′-to-3′ direction along a backbone chain, the close contacts are in the major groove; for a pyrimidine-to-purine step the clash is in the minor groove [*see illustration on page 110*]. How might these close contacts between the edges of purine bases be relieved?

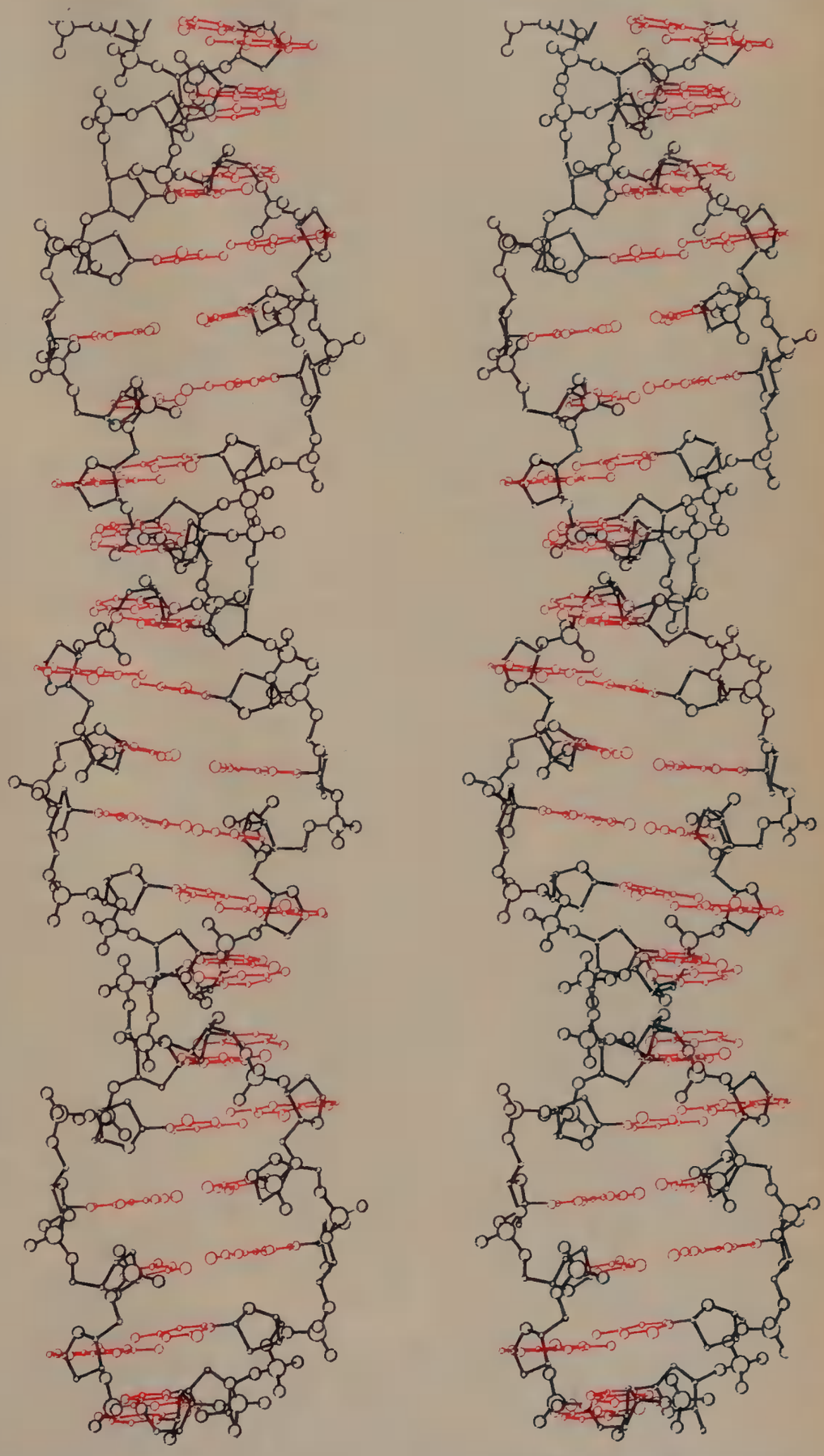
One way is simply to open the base-roll angle on the side where the close contacts take place. For the minor-groove clash of a pyrimidine-purine step the proper response would be to open the angle between the mean plane of the base pairs toward the minor groove, making the roll angle more positive. At purine-pyrimidine steps the roll angle should be opened toward the major groove, making the angle more negative. Drew and I had observed this sequence-dependent behavior of roll angle in the *CGCGAATTCGCG* structure, but our explanation was more convoluted and less plausible than Calladine's.

No matter on which side of a base pair the clash with a neighbor takes place, it always can be lessened by decreasing the local helical-twist angle between the two base pairs. This suggests that helical-twist variation might have the same simple stereochemical explanation as roll angle has. A purine-purine clash across the helix also can be relieved directly by damping down the propeller twist in one or both base pairs, or indirectly by sliding a base pair along its long axis so that the purine is partially removed from the helix stack. Calladine concluded from his analysis of the latter effect that steric clash in the minor groove is roughly twice as pronounced as the clash in the major groove and therefore requires twice the adjustment in local helix parameters.

Predicting the Variation

Starting from Calladine's premises, our group (now working at the University of California at Los Angeles) has developed four simple sum functions by which to predict, from a DNA's base sequence, the local variations in helical twist, in roll angle, in propeller twist and also in torsion angle (a measure of the conformation of the backbone chain around the C4′–C3′ bond). The functions for propeller twist and torsion angle seem to apply only to *B* DNA, but the first two functions are valid for both *B* and *A* helixes. As an example of the prediction process, the sum function Σ_1 for helical-twist angle is generated as follows.

1. At each interval from one base pair to the next, assign a number that reflects the relative tendency at that step toward a reduction in helical twist in order to



STEREO PAIR of the extended *Z* helix is a reiteration of *GCGC*. Note that the structure of *Z* DNA and that of *A* DNA are in many ways the inverse of each other. The *Z* helix is tall and slim, with a deep minor groove, a flattened major groove and small propeller twist. The *A* helix is short and broad, with a deep major groove, a very shallow minor groove and large propeller twist.


Panasonic
Portable Video Cassette Recorder
VHS
STEREO
Panasonic
Programmable Tuner/Timer

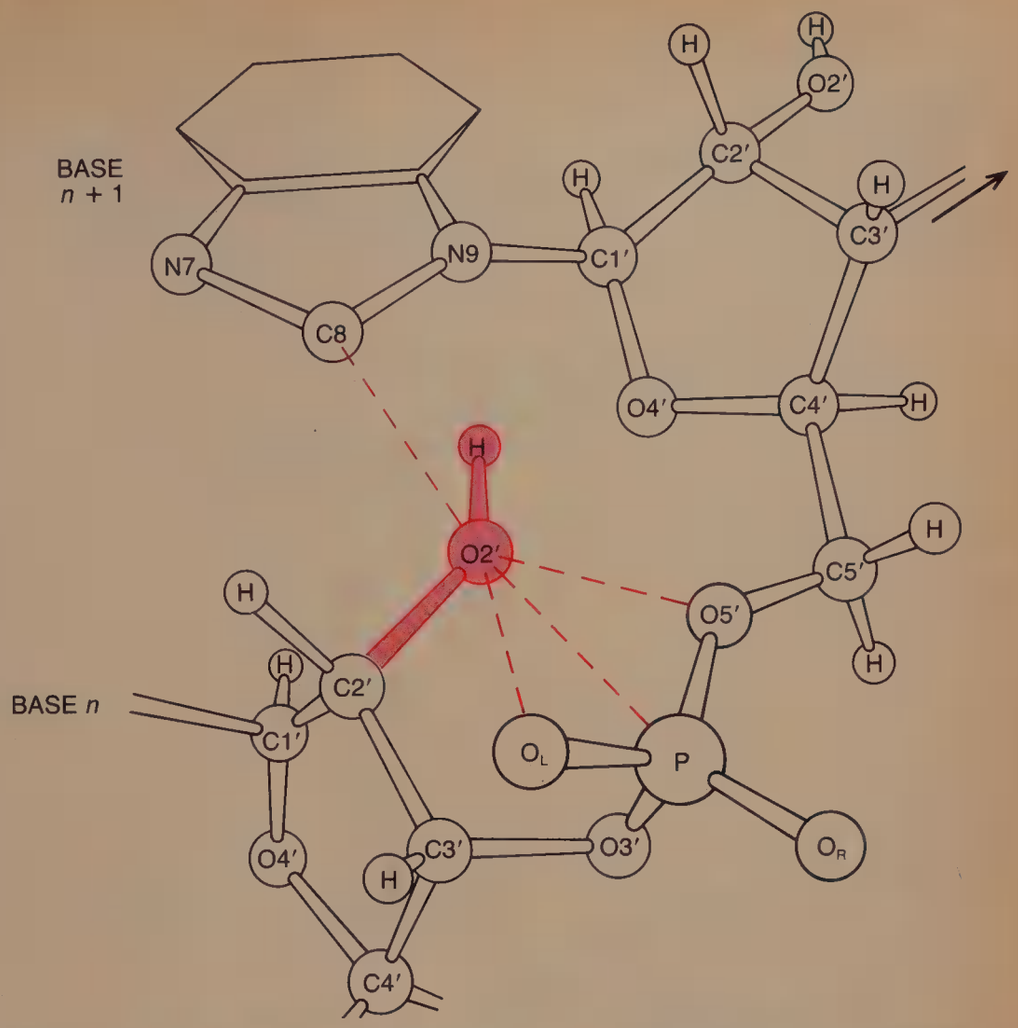
lessen stereochemical clash. Let this number be −2 for purine-pyrimidine steps with their major-groove clash, let it be −4 for pyrimidine-purine steps with their more pronounced minor-groove contact and let it be zero for purine-purine and pyrimidine-pyrimidine steps, which lack steric hindrance.

2. Assign numbers of opposite sign but half the magnitude to the flanking steps, since closing down a central interval by rotating both base pairs will open up the intervals on both sides by half as much. Hence the total contribution to the sum function at each purine-pyrimidine step and its two flanking steps will be +1, −2, +1, and the values for each pyrimidine-purine step and its neighbors will be +2, −4, +2.

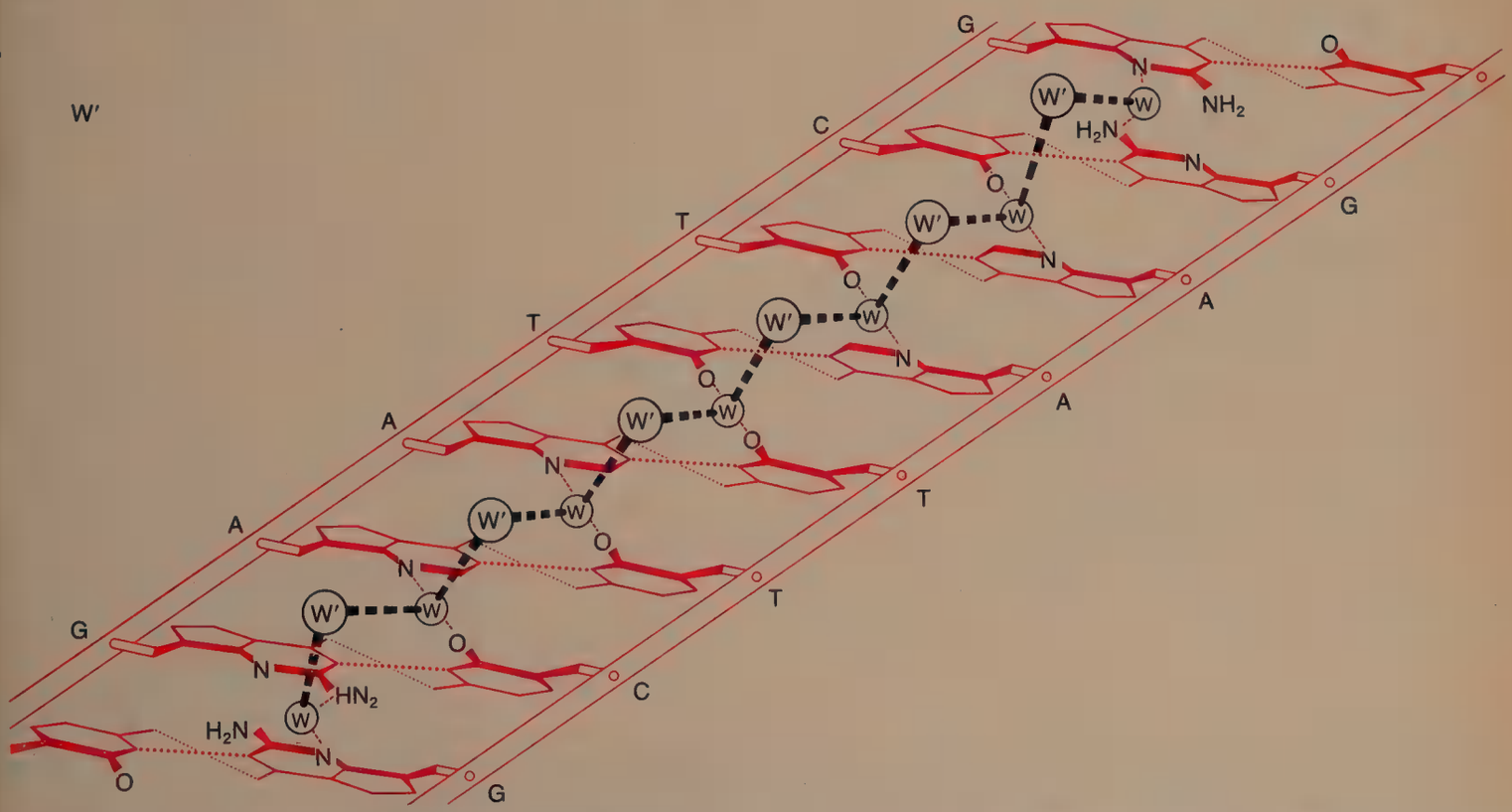
3. Add the contributions from all steps to obtain the full sum function, Σ_1, which measures the local sequence-induced variation from the average helical-twist angle.

The derivation of sum function Σ_1 is illustrated for the *B*-helical dodecamer *CGCGAATTCGCG,* and the predicted and observed values are compared, in the illustration on page 111. A linear-regression comparison of the observed angles with Σ_1 yields a correlation coefficient of .994, or a highly significant correlation between the two quantities. The sum function faithfully reproduces the ups and downs of the observed angles except at the two ends of the sequence, where intermolecular overlap in the crystal distorts the helix.

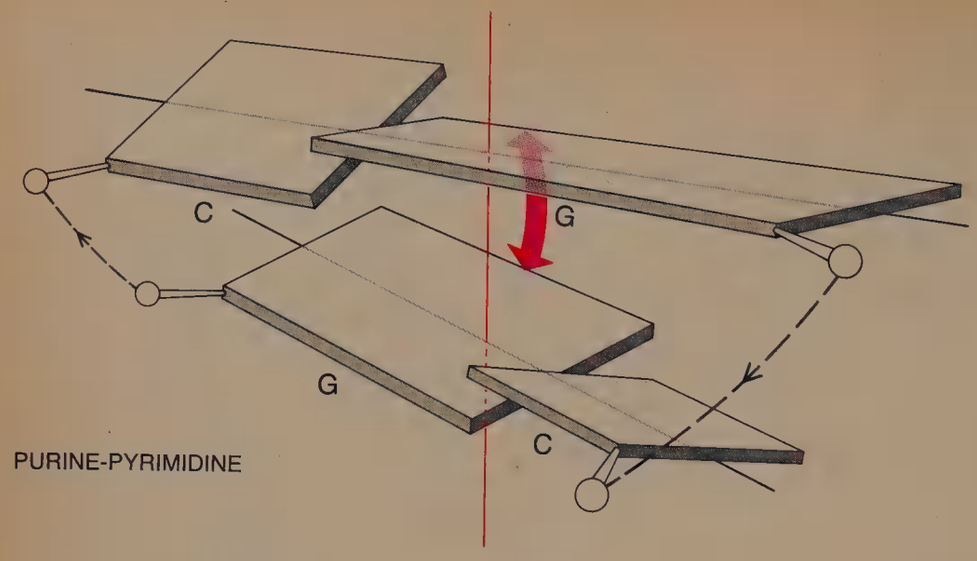
A similar procedure can be followed

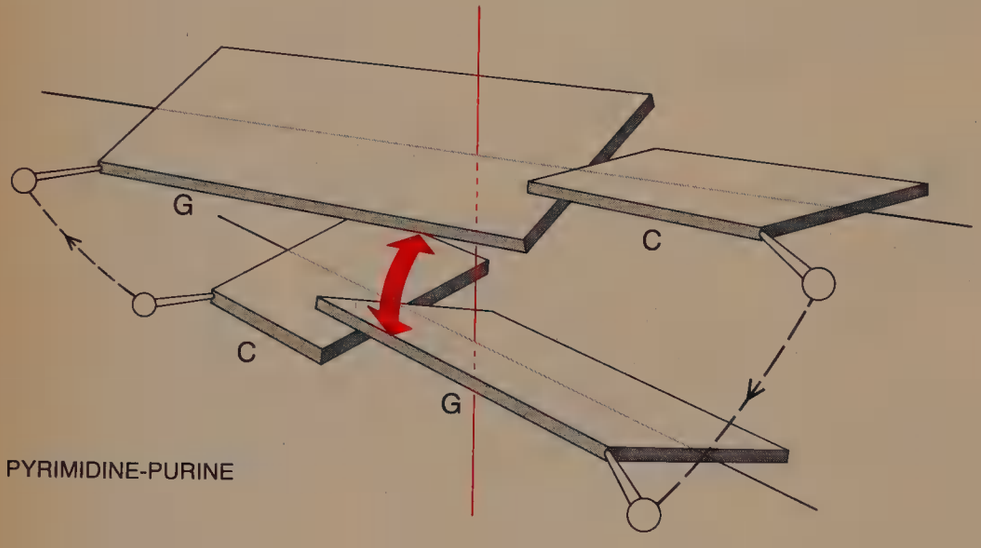


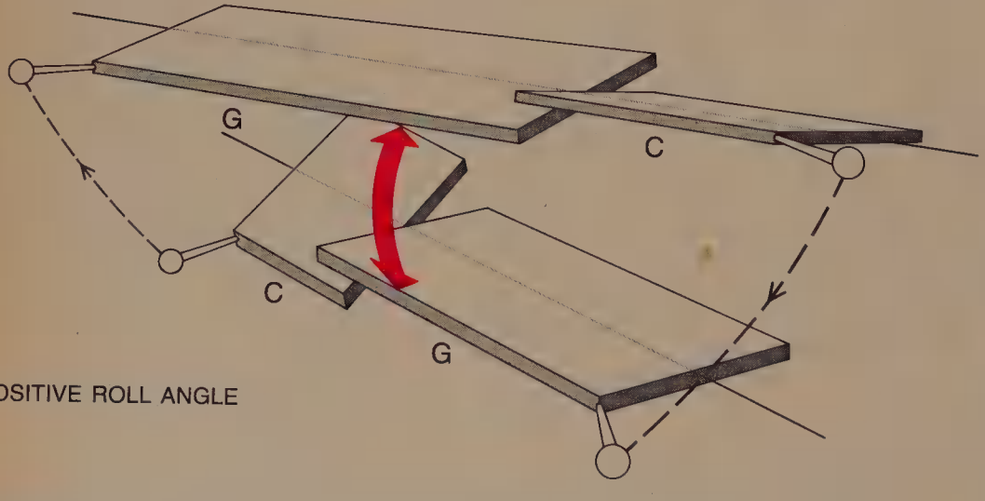
RNA does not adopt the *B*-helical structure when it forms a self-complementary double helix. The reason seems to be that RNA has ribose sugar rings, with a hydroxyl group (OH) where the deoxyribose ring of DNA has only a hydrogen atom. Here an OH has been added to a *B*-DNA backbone to simulate a *B* RNA. There is an uncomfortably close contact between the added oxygen atom and four atoms of the adjacent phosphate, sugar and base (*broken lines*).



SPINE OF HYDRATION (*black*) in *B* DNA is formed by a zigzag array of water molecules. This is a view into the minor groove of the unrolled helix of the *B*-DNA dodecamer. The backbones are highly schematized and propeller twist is exaggerated. A bottom layer of water molecules (*W*) is hydrogen-bonded to nitrogen and oxygen atoms on adjacent base pairs; these molecules are linked by a second layer of water molecules (*W′*) to build the spine. The spine keeps the minor groove from opening up to form the shallow depression seen in *A* DNA, and it is therefore thought to be a major factor in making *B* DNA the stable helical structure under conditions of high hydration.







PROPELLER TWIST affects contacts between purines on opposite strands at adjacent base pairs. Here purines and pyrimidines are schematized as planks, the C1′ atoms of sugar rings are designated by small spheres and the backbones are reduced to broken lines, with arrows pointing in the 5′-to-3′ direction. For a purine-pyrimidine step (*top*) positive propeller twist brings the edges of two purines too close together in the major groove (*colored arrow*). For ■ pyrimidine-purine step (*middle*) the unacceptably close contact is in the minor groove. One strategy (*bottom*) for relieving this clash is to open up the base-roll angle (formed by the tilt of adjacent base pairs as a whole) where the clash takes place, in this case in the minor groove.

to calculate sum function Σ_2, which measures the local variation in roll angle, except that the corrections at pyrimidine-purine steps have reversed signs $(-2, +4, -2)$ because at such steps steric hindrance is relieved by opening the angle on the minor-groove side, making the angle more positive. The derivation of Σ_2 and its comparison with the observed roll angles are illustrated for *CGCGAATTCGCG*. Again the increase and decrease of the roll angles is predicted accurately except at the extreme ends of the helix.

If these two sum functions worked for only one helix of one type, the method would be like many scientific bright ideas that work once in the hands of their inventor but never again for anyone else. It turns out, however, that when the same rules are applied to the four known *A* helixes, the results are fully as good as they are for the *B* sequence. One of the *A* structures offered a particularly stringent test. The two tetramer molecules of *CCGG*/*CCGG* stack on top of each other in the crystal to form what amounts to an octamer even though the central two base pairs are not bridged by phosphates. In spite of this lack of covalent connection, both the helical-twist angle and the base-plane roll angle across the central step are exactly what would have been predicted by the sum functions for an intact, covalent octamer. Stacking forces between base pairs thus appear to dominate over the influence of the sugar-phosphate backbone in determining double-helical DNA structure. Both the twist function and the roll function also hold for *A*-helical *GGTATACC*, and only the twist function begins to fail in the case of the RNA-DNA hybrid base pairs in the mixed helix (*GCG*)*TATACGC*.

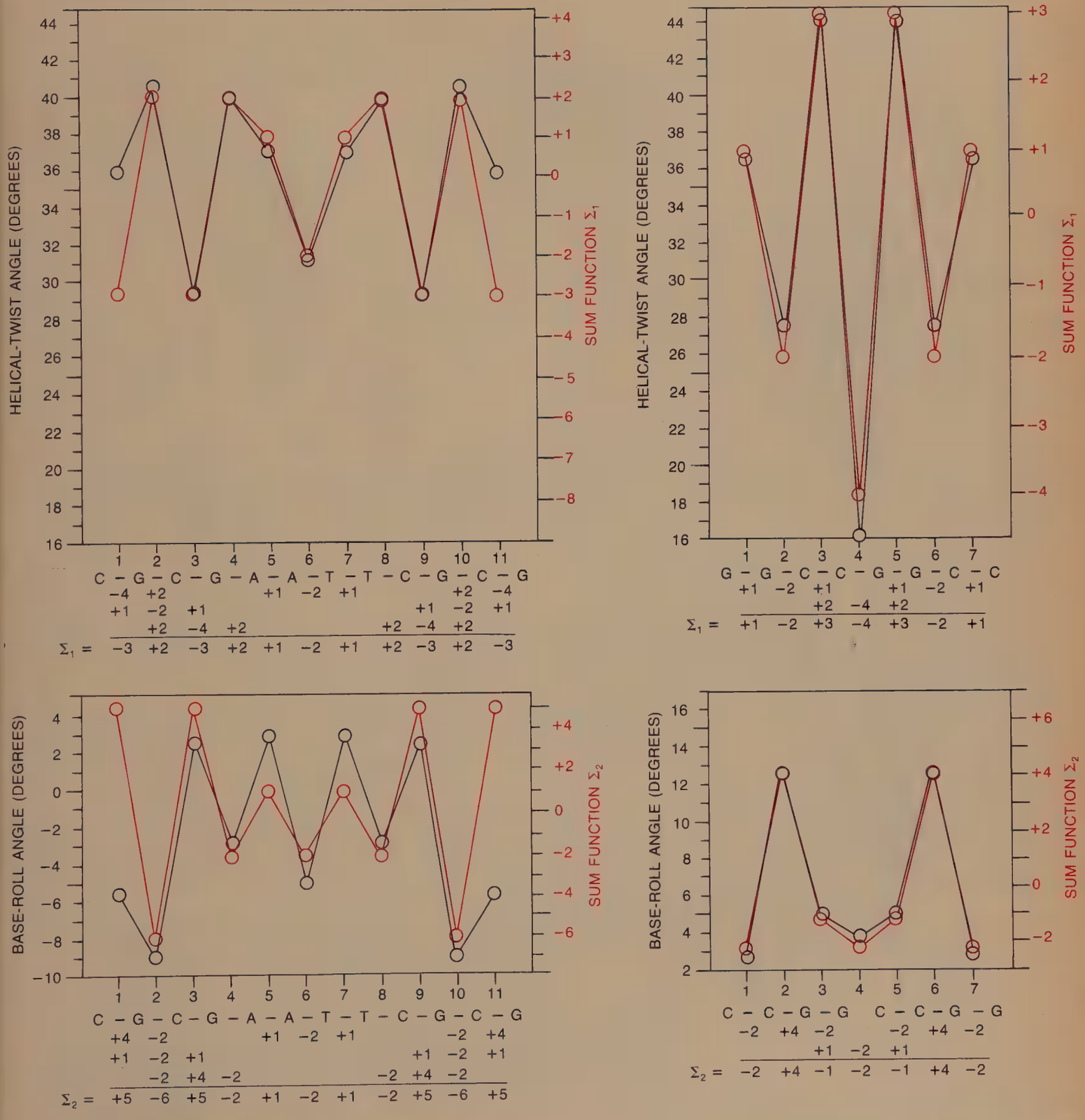
From Structure to Control

Do these local fluctuations in helix structure actually influence the reading of DNA base sequences by repressors and similar control proteins? If one accepts the idea that the readout process involves hydrogen-bond formation between protein side chains and nitrogen and oxygen atoms on the edges of the bases, it is hard to imagine that these variations of as much as 15 degrees from the mean would not have a major effect on the recognition process. Perhaps the sequence-induced fluctuations represent fine tuning of the fit between DNA and protein, adding to the readout process a dimension that goes beyond the mere arrangement of hydrogen-bond donors and acceptors along the edge of each base pair.

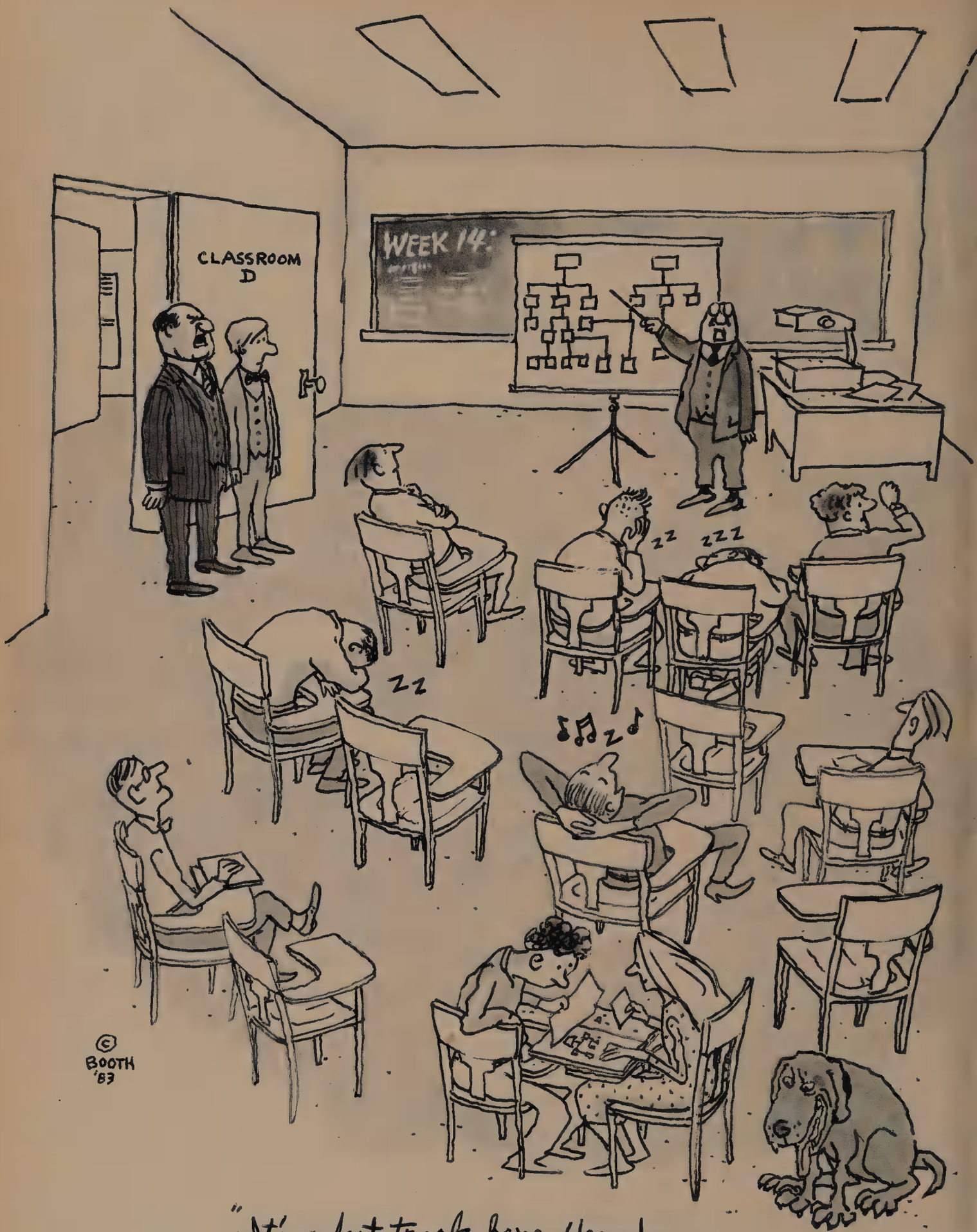
Whether that is the case can only be established by X-ray crystal-structure analysis of a recognition protein in a complex with its particular DNA sequence. Several studies of such com-

plexes are in progress, involving the *lac, lambda* and *cro* repressors, catabolite-activator protein (CAP) and various restriction enzymes. John Rosenberg of the University of Pittsburgh has recently calculated an electron-density map of the *Eco RI* restriction enzyme bound to its recognition site on the dodecamer *CGCGAATTCGCG*. He reports that the bound DNA remains in the *B* form but that the helix is distorted.

The questions one can sensibly ask are always influenced by what one already knows. A new advance in knowledge can remove an issue from the realm of idle speculation and place it at the center of scientific inquiry. This is what the recent single-crystal X-ray analyses have done for such questions as: How does base sequence in double-helical DNA affect the structure and behavior of the helix? To what extent is this effect significant in the readout of information? The answers are only beginning to come in, but clearly the questions are well worth asking and will lead eventually to better understanding of genetic control.



LOCAL VARIATIONS in helical-twist angle and base-roll angle are predicted from the base sequence by the sum functions Σ_1 and Σ_2. Here the method is demonstrated for the *B*-helical dodecamer (*top left and bottom left*) and for an *A*-helical octamer (*top right*) and pseudo-octamer (*bottom right*). In each case observed values are shown by the black curves and the scales at the left, predictions by the colored curves and the scales at the right. The derivation of the sum function for each step between adjacent base pairs is shown below each graph. The curves for the observed values and the predicted values are quite congruent: the correlation coefficients between the two sets of values are .994 for the *B*-helix twist angle, .917 for the *B*-helix roll angle, .991 for the *A*-octamer twist angle and .995 for the *A*-pseudo-octamer roll angle. The pseudo-octamer (actually two stacked tetramers, unbridged by phosphates) behaves in this respect like a true octamer; base-pair stacking seems to be more important than the connecting phosphate backbone is for local helix structure.



"It's a fast track here, Herndon.
Once you get through our training program."

These days, the best-run companies in America are moving faster than ever. Just about everywhere you look, you'll see all sorts of powerful new ideas. Entrepreneurial management styles. Personal computers. Teleconferencing.

For some reason though, the training department has been left a bit behind.

But now, thanks to a company called Wicat, the technology of training has finally come into its own.

HOW TO CUT TRAINING TIME AND COST IN HALF.

Simple. Use the power of computers to provide intense, individualized study for each trainee.

Some of the world's largest corporations put a Wicat learning station in front of each trainee. And train them better, in less than half the time. And at less than half the cost.

DEVELOP COURSES OVERNIGHT; NO PROGRAMMING NEEDED.

Of course, there are other computer-based training systems. But they demand sophisticated programming skills. Wicat doesn't.

For example, it took a skilled programmer six weeks to develop one course on PLATO™ (the other well-known training system). To develop the same course on the Wicat system, it took a non-programmer less than two days.

LASERDISCS VS. OVERHEAD PROJECTORS.

There's nothing like realism to make training stick. Which is why Wicat was the first to incorporate the extraordinary graphics, sound, and simulation capability of the laserdisc into computer-based training.

Imagine, if you will, 30 trainees sitting through a lecture on shop floor safety. Now imagine each trainee in a one-on-one laserdisc simulation of machine shop hazards. Where they can learn from their mistakes. And live through them.

THE TOTAL TRAINING SOLUTION.

At Wicat, we design and manufacture our own hardware. In fact, we spent four years in development, to make this the most productive training machine possible.

Our WISE authoring system sets new standards for ease of course development. And our large staff is ready to design custom courseware for any applications.

We even offer software that handles everything from scheduling to progress reporting to testing.

Call (801) 224-6400 for details. Or write to Wicat Systems, P.O. Box 539, 1875 South State Street, Orem, UT 84057.

And let us help you get your training program on the fast track.

WICATsystems™

PLATO is a trademark of Control Data Corp.

Quantum Gravity

In a quantum-mechanical theory of gravitation the very geometry of space and time would be subject to continual fluctuations, and even the distinction between past and future might become blurred

by Bryce S. DeWitt

Among the forces of nature gravity seems to have a special status. Other forces, such as electromagnetism, act *in* spacetime, which serves merely as the setting for physical events. Gravity is quite different. It is not a force imposed on the passive background of space and time; rather it constitutes a distortion of spacetime itself. A gravitational field is a "curvature" of spacetime. This is the conception of gravity achieved by Einstein in what he described as the hardest work of his life.

The qualitative distinction between gravity and other forces becomes all the more apparent when one tries to formulate a theory of gravitation consistent with the precepts of quantum mechanics. The quantum world is never still. In the quantum field theory of electromagnetism, for example, the value of the electromagnetic field is continually fluctuating. In a universe governed by quantum gravity the curvature of spacetime and even its very structure would be subject to fluctuations. Indeed, it is possible that the sequence of events in the world and the meaning of past and future would be susceptible to change.

One might suppose that if such phenomena existed, they would surely have been noticed by now. As it happens, however, any explicitly quantum-mechanical effects of gravitation would be confined to an exceedingly small scale, a scale to which attention was first drawn by Max Planck in 1899. In that year Planck introduced his famous constant, called the quantum of action and denoted $\hbar$. He was trying to make sense of the spectrum of black-body radiation, the light that escapes by way of a small opening from a hot cavity. In a curious aside he noted that his constant, when it is combined with the velocity of light and Newton's constant of gravitation, establishes an absolute system of units. These units set the scale for quantum gravity.

Planck's units are totally unrelated to everyday physics. His unit of length, for example, is 1.61×10^{-33} centimeter. This is 21 powers of 10 smaller than the diameter of an atomic nucleus. It bears roughly the same relation to nuclear dimensions as the size of a human being bears to that of our galaxy. The Planck unit of time is even more fantastic: 5.36×10^{-44} second. To probe these scales of distance and time experimentally, using instruments built with present technology, one would need a particle accelerator the size of the galaxy!

Because experiment can offer no guidance, quantum gravity is unusually speculative. Nevertheless, it is basically conservative in spirit. It takes well-established existing theory and merely pushes it to its extreme logical conclusions. In its barest essentials quantum gravity aims to combine three component theories: special relativity, the Einstein theory of gravitation and quantum mechanics, nothing more. Such a synthesis has not yet been fully achieved, but much has been learned from the effort. Furthermore, the development of a working theory of quantum gravity offers the only known path toward understanding the origin of the big bang and the ultimate fate of black holes, events that can be taken as marking the beginning and the end of the universe.

Of the three component theories of quantum gravity special relativity came first historically. It is the theory that unites space and time through the (experimentally confirmed) postulate that the velocity of light is the same for all observers moving in regions of empty space free of external forces. The consequences of this postulate, which was introduced by Einstein in 1905, can be described with the aid of a spacetime diagram, a graph bearing curves that represent the positions of objects in space as a function of time. The curves are called world lines.

For the sake of simplicity I shall ignore two of the spatial dimensions. A world line can then be drawn on a two-dimensional graph where spatial distances are measured horizontally and time intervals are measured vertically. A vertical straight line is the world line of an object at rest in the frame of reference chosen for the measurement. A tilted straight line is the world line of an object moving at a constant velocity in the chosen frame of reference. A curved world line represents an object undergoing acceleration.

A point on the spacetime diagram specifies both a position in space and a moment in time and is called an event. The spatial distance between two events depends on the frame chosen, and so does the time interval. Indeed, the concept of simultaneity is frame-dependent. Two events that are connected by a horizontal line in a chosen frame are simultaneous in that frame but not in other frames.

To establish a relation between frames in relative motion one must introduce a common unit for the measurement of space and time. The velocity of light serves as a conversion factor, relating a given distance to the time needed for light to traverse it. Here I shall adopt the meter as the unit of both space and time. One meter of time is about $3\frac{1}{3}$ nanoseconds (billionths of a second).

When space and time are measured in the same units, the world line of a photon (a quantum of light) is tilted at 45 degrees. The world line of any material object has a tilt from the vertical that is always less than 45 degrees, which is another way of saying that its velocity is always less than that of light. If the world line of any object or signal were tilted at more than 45 degrees from the vertical, the object or signal would seem to certain observers to be moving backward in time. By setting up a relay of faster-than-light signals a person could transmit information into his own past, thus violating the principle of causality. Such signals are outlawed by provisions of special relativity.

Consider two events on the world line of an unaccelerated observer. Suppose the events, in a particular frame of reference, are separated by four meters of space and five meters of time. Then in that frame the observer is moving at four-fifths the speed of light. In another

frame his speed would be different, and so would the associated space and time intervals. There is one quantity, however, that would be the same in all frames of reference. This invariant quantity is called the "proper time" between the two events; it is the time interval measured by a clock the observer carries with him.

In the chosen frame of reference the world line between the two events is the hypotenuse of a right triangle having a base of four meters and a height of five meters. The "proper time" corresponds to the "length" of this hypotenuse, but it is calculated in an unusual way: by means of a "pseudo-Pythagorean theorem." As with the ordinary Pythagorean theorem, one first squares the sides of the triangle. In special relativity, however, the square of the hypotenuse is equal to the difference of the squares of the sides rather than to the sum.

In the present example the proper time is three meters. It remains three meters in the frame of reference of any unaccelerated observer. The invariance of the proper time is what unites space and time into the single entity spacetime. The geometry of spacetime, being based on a pseudo-Pythagorean theorem, is not Euclidean, but in many ways it is analogous. In Euclidean geometry, among all the lines joining two points a straight line can be defined as one whose length is an extremum. The same is true in the geometry of spacetime. In Euclidean geometry, however, the extremum is always a minimum, whereas in spacetime it is a maximum whenever the two points can be connected by a world line requiring no faster-than-light travel.

In 1854 the German mathematician G. F. B. Riemann invented a generalization of Euclidean geometry for curved spaces. Two-dimensional curved spaces had been studied since antiquity. They are called curved surfaces and are usually viewed from the perspective of ordinary three-dimensional Euclidean space. Riemann showed that a curved space can have any number of dimensions and that it can be studied intrinsically. It does not need to be imagined as

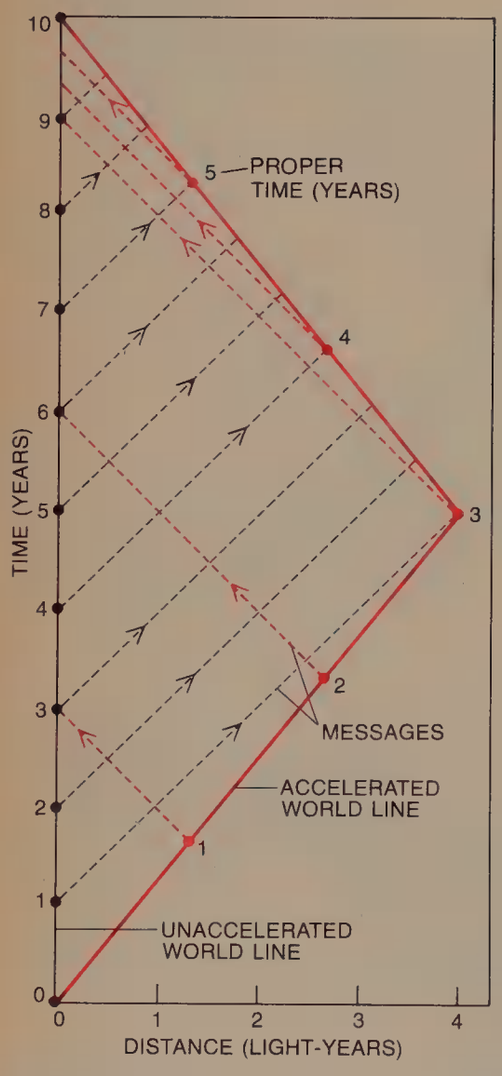


LIGHT CONE, which defines the regions of the universe accessible from a given point in space and moment in time, would become an ill-defined concept in a theory of quantum gravity. The cone (*a*) is a surface in four-dimensional spacetime, but it is drawn here with one spatial dimension suppressed. If gravitation is quantized, the shape of the cone can fluctuate wildly over short distances (*b*). Actually the fluctuations could not be perceived directly; instead the light cone would act as if it were fuzzy. As a result the question of whether two points in spacetime could communicate with each other (by means of signals moving no faster than the speed of light) could be given only a probabilistic answer (*c*).

embedded in a Euclidean space of higher dimension.

Riemann also pointed out that the physical space we live in may be curved. In his view the question could be decided only by experiment. How might one carry out such an experiment, at least in principle? Euclidean space is said to be flat. A flat space has the property that parallel straight lines can be drawn to create a uniform rectangular grid. What would happen if one attempted to draw such a grid on the surface of the earth in the belief that the earth is flat?

The result can be seen from an airplane on any clear day over the cultivated regions of the Great Plains. The land is subdivided by east-west and north-south roads into square-mile sections. The east-west roads often extend in unbroken lines for many miles, but not the north-south roads. Following a road northward, there are abrupt jogs to the east or west every few miles. The jogs are forced by the curvature of the earth. If the jogs were eliminated, the roads would crowd together, creating sections of less than a square mile.



WORLD LINE defines a path through space and time. Here two world lines are shown in a version of Einstein's twins paradox. The "bent" world line of the twin who undergoes acceleration at the turnaround point in his journey appears to be the longer one, but that twin records the shorter "proper time." Indeed, a straight line represents the longest interval between two points in a spacetime diagram. The emission and arrival times of the messages transmitted by the twins are shown.

In the three-dimensional case one can imagine building a giant scaffold in space out of straight rods of equal length joined at angles of precisely 90 degrees and 180 degrees. If space is flat, the construction of the scaffold would proceed without difficulty. If space is curved, one would eventually have to begin shortening the rods or stretching them to make them fit.

The same generalization Riemann introduced for Euclidean geometry can be applied to the geometry of special relativity. The generalization was carried out by Einstein between 1912 and 1915 with the help of the mathematician Marcel H. Grossmann. The result is a theory of curved spacetime. In Einstein's hands it was developed into a theory of gravitation. In special relativity gravitational fields are assumed to be absent and spacetime is flat. In a curved spacetime a gravitational field is present; indeed, "curvature" and "gravitational field" are synonymous.

Because Einstein's gravitational field theory is a generalization of special relativity, he called it general relativity. The name is a misnomer. General relativity is actually less relativistic than special relativity. The complete featurelessness of flat spacetime, its homogeneity and isotropy, are what ensure that positions and velocities are strictly relative. As soon as spacetime acquires "bumps," or local regions of curvature, it becomes absolute because position and velocity can be specified with respect to the bumps. Spacetime, instead of being merely a featureless arena for physics, is itself endowed with physical properties.

In Einstein's theory curvature is produced by matter. The relation between the amount of matter and the degree of curvature is simple in principle but complicated to calculate. Twenty functions of the coordinates of a point in spacetime are needed to describe the curvature at that point. Ten of the functions correspond to a part of the curvature that propagates freely in the form of gravitational waves, or "ripples of curvature." The other 10 functions are determined by the distribution of mass, energy, momentum, angular momentum and internal stresses in the matter, as well as by Newton's gravitational constant, G.

With reference to the mass densities encountered on the earth G is a very small constant. It takes a lot of mass to bend spacetime appreciably. The reciprocal quantity $1/G$ can be regarded as a measure of the "stiffness" of spacetime. In terms of everyday experience spacetime is very stiff. The entire mass of the earth induces a spacetime curvature that is only about a billionth the curvature of the earth's surface.

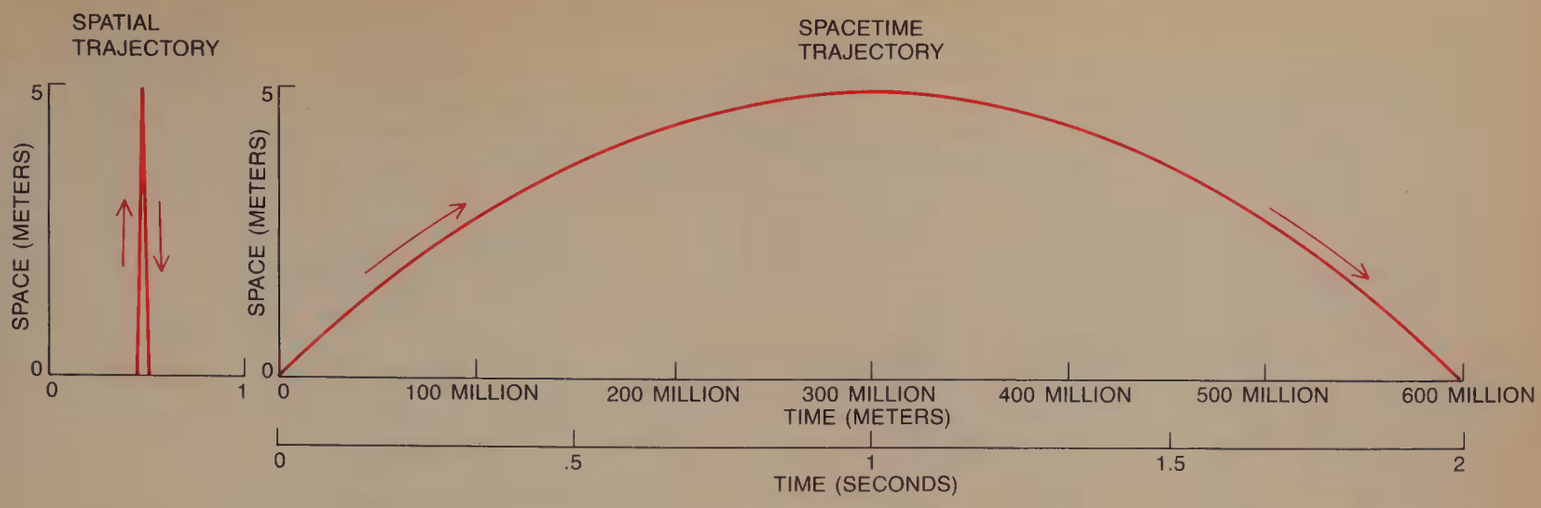
In Einstein's theory a freely falling or freely orbiting body follows a geodesic world line. A geodesic connecting two spacetime points is a world line of extremal length between them; it is a generalization of the concept of a straight line. If a curved spacetime is imagined as embedded in a flat space of higher dimension, a geodesic appears as a curved line.

The effect of curvature on a moving body has often been illustrated by a model in which a ball rolls on a distorted rubber sheet. The model is misleading in that it can represent only spatial curvature. In real life we are constrained to remain in the four-dimensional universe of space and time. Moreover, we cannot avoid motion in this universe, because we are hurtling forward in time. Time is the key element. It turns out that although space is curved in a gravitational field, curvature in time is much more important. The reason is the high value of the velocity of light, which is the magnitude relating the scale of space to that of time.

Near the earth the curvature of space is so slight that it cannot be detected by static measurements. Yet our headlong rush in time is so fast that in dynamical situations the curvature becomes noticeable, much as a slight bump on a speedway, which would pass unnoticed by a pedestrian, becomes a hazard to a fast-moving automobile. Although space near the earth appears flat to a high degree of precision, we can see the curvature of space*time* by merely throwing a ball into the air. If the ball is in the air for two seconds, it follows an arc with a height of five meters. Light travels 600,000 kilometers in two seconds. If one imagines the five-meter-high arc stretched horizontally to a length of 600,000 kilometers, the curvature of the arc is the curvature of spacetime.

Riemann's introduction of the idea of curved spaces initiated another rich area of mathematical study: topology. Boundaryless two-dimensional surfaces had been known to exist in an infinite variety of types that cannot be continuously deformed into one another; a sphere and a torus are two simple examples. Riemann pointed out that the same is true of higher-dimensional curved spaces, and he took some initial steps toward their classification.

Curved spacetimes (or, more precisely, models of curved spacetime) also exist in an infinite variety of topological types. As candidates for a description of the real universe some of the models can be rejected because they lead to paradoxes of causality or because in them known physical laws cannot be made to hold. There still remain a staggering number of possibilities.



CURVATURE OF SPACETIME in the presence of mass constitutes a gravitational field. When a ball is thrown five meters into the air (*left*), it stays aloft for two seconds. Its up-and-down motion reveals the curvature of spacetime near the surface of the earth. Although the curvature of the trajectory is readily visible, it is actually very small when space and time are measured in the same units. For example, seconds can be converted into meters by multiplying them by the speed of light, or 300 million meters per second. When that is done, the trajectory becomes an exceedingly shallow arc 600 million meters long and five meters high (*right*). The height is exaggerated.

One notable model of the universe was proposed by the Russian mathematician Alexander A. Friedmann in 1922. In special relativity spacetime is viewed as being not only flat but also infinite in extent in both space and time. In Friedmann's model every three-dimensional spatial cross section of spacetime is finite in volume and has the topology of a 3-sphere, a space that can be embedded in a four-dimensional Euclidean space in such a way that all its points are equidistant from a given point. The model has been a favorite of cosmologists ever since the expansion of the universe was discovered by Edwin P. Hubble in the 1920's. When Friedmann's model is combined with Einstein's theory of gravitation, it predicts a big bang at an initial moment of infinite compression, followed by an expansion that slows down over billions of years because of the mutual gravitational attraction of all the matter in the universe.

A Friedmann spacetime has the property that every closed curve drawn in it can be continuously shrunk to a point. Such a spacetime is said to be simply connected. The real universe may not have this property. The Friedmann model seems to describe very well the region of space within a few billion light-years of our own galaxy, but we cannot see the entire universe.

A simple example of a multiply connected universe is one whose structure is repeated ad infinitum, like a wallpaper pattern, in a given spatial direction. Every galaxy in such a universe is a member of an infinite series of identical galaxies separated by some fixed (and necessarily enormous) distance. If the members of a series are truly identical, it is questionable whether they should be considered distinct. It is more economical to view each series as representing just one galaxy. Hence a journey from one member of the series to the next returns a traveler to his starting point, and a line tracing such a journey is a closed curve that cannot be shrunk to a point. It is like a closed curve on the surface of a cylinder that goes around the cylinder once. The repeating universe is called a cylindrical universe.

Another example of a multiply connected structure, on a far smaller scale, is the "wormhole," introduced in 1957 by John Archibald Wheeler, now of the University of Texas at Austin. A two-dimensional wormhole can be constructed by cutting two circular holes in a two-dimensional surface and smoothly joining the cut edges [*see illustration on page 118*]. The procedure is the same in three dimensions; it is just harder to visualize.

Because the two holes can be at a great distance in the original space and yet close together through the "throat" connecting them, the wormhole has become a popular device in science fiction for getting from one place to another much faster than light can get there: just punch two holes in space, connect them and crawl through the throat. Unfortunately even if one could construct a hole puncher (which is doubtful), the scheme would not work. If the geometry of spacetime is governed by Einstein's equations, the wormhole is a dynamical object. It turns out that the two holes it joins are necessarily black holes and anything that gets into them can never get out again. What happens is that the throat "pinches off," and everything inside is crushed to infinite density before it can get to the other side.

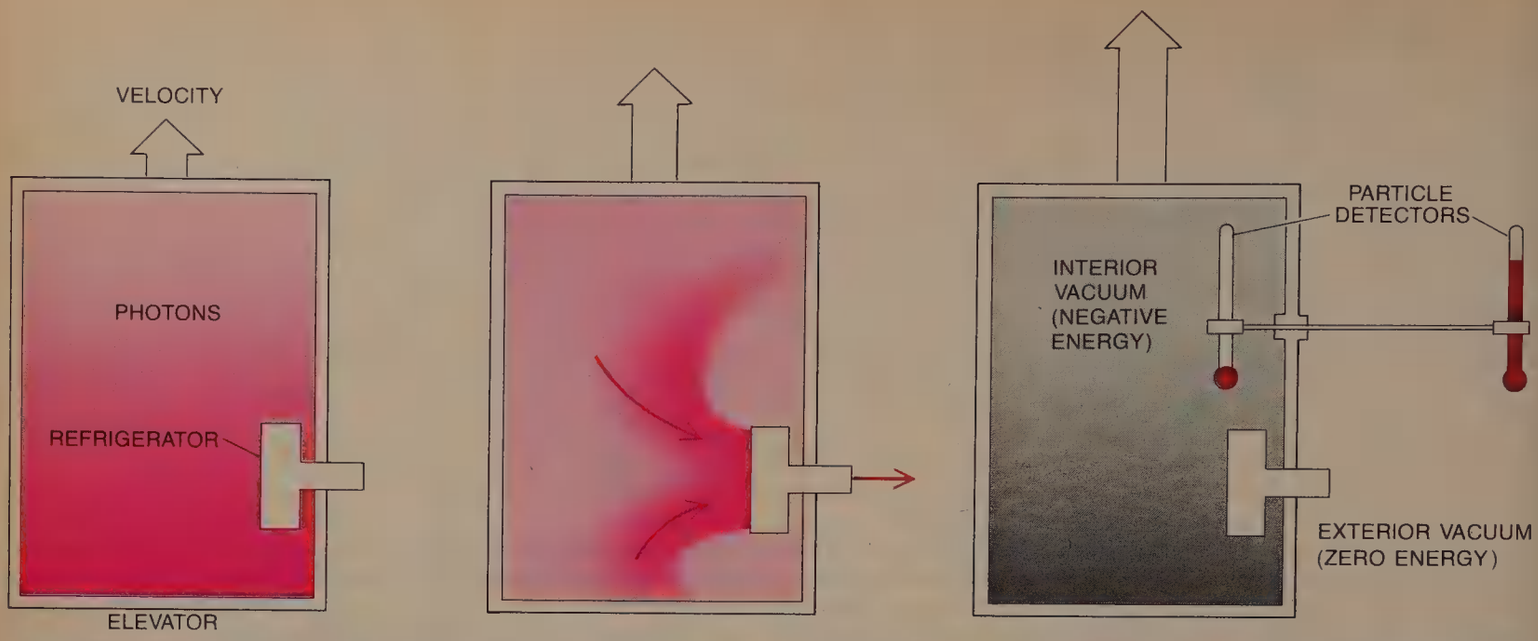
Quantum mechanics, the third component of quantum gravity, was invented in 1925 by Werner Heisenberg and Erwin Schrödinger, but their initial formulation took no account of the theory of relativity. Its success was nonetheless immediate and brilliant, for there lay waiting to be explained a host of experimental observations in which quantum effects dominate and relativity plays a minor or negligible role. It was known, however, that the electrons in some atoms reach speeds that are a sizable fraction of the speed of light, and so the search for a relativistic quantum theory was not long delayed.

By the mid-1930's it was fully understood that when the quantum theory is combined with relativity, a number of entirely new facts can be deduced. The two most fundamental facts are the following. First, every particle is associated with a type of field and every field is associated with a class of indistinguishable particles. No longer could the electromagnetic and gravitational fields be considered the only basic fields of nature. Second, there are two types of particles classified according to their (quantized) spin angular momentum. Those with spins $\frac{1}{2}\hbar$, $\frac{3}{2}\hbar$ and so on obey the exclusion principle (no two in the same quantum state); those with spins 0, $\hbar$, $2\hbar$ and so on are gregarious.

These astonishing consequences of uniting special relativity and quantum mechanics have been confirmed repeatedly in the past half century. Relativity and quantum theory together yield a theory that is greater than the sum of its parts. The synergistic effect is even more pronounced when gravity is included.

In classical physics flat, empty spacetime is called the vacuum. The classical vacuum is featureless. In quantum physics a much more complex entity, with a rich structure, is given the name vacuum. Its structure arises from the existence in it of nonvanishing free fields, that is, fields far from their sources.

A free electromagnetic field is mathematically equivalent to an infinite collection of harmonic oscillators, which can be represented as springs with attached masses. In the vacuum every oscillator is in its ground state, or state of



ACCELERATING ELEVATOR CAR is apparatus for a thought experiment that bears on the nature of the vacuum in quantum mechanics and on the effect either acceleration or gravitation has on the vacuum. The car is assumed to be empty and sealed, so that initially there is a perfect vacuum both inside the car and outside. When the acceleration begins, however, an electromagnetic wave is emitted by the floor, and the car becomes filled with a dilute gas of photons, or quanta of electromagnetic radiation (*left*). A refrigerator driven by some outside source of energy pumps out the photons (*middle*). When all the photons have been removed, photon detectors measure the energy of the vacuum both inside and outside (*right*). Because the instrument outside is accelerating through the vacuum, it responds to quantum-mechanical fluctuations of fields that permeate space even in the absence of particles. The detector inside is at rest with respect to the car and does not sense the fluctuations. It follows that the vacua inside and outside the car are not equivalent. If the "standard" vacuum outside the car is defined as having zero energy, the vacuum inside must have negative energy. The photons removed by the refrigerator would have to be restored in order to bring the energy up to zero. A gravitational field can also create a vacuum with negative energy.

lowest energy. When a classical (nonquantum-mechanical) oscillator is in its ground state, it is motionless at a well-defined point. This is not true of a quantum oscillator. If a quantum oscillator were at a definite point, its position would be known with infinite precision; the uncertainty principle then implies that it would have infinite momentum and energy, which is impossible. In the ground state of a quantum oscillator neither position nor momentum is precisely fixed. Both are subject to random fluctuations. In the quantum vacuum it is the electromagnetic field (and every other field) that fluctuates.

Although the field fluctuations in the quantum vacuum are random, they are of a special kind. They satisfy the principle of relativity in that they "look" the same to every unaccelerated observer, no matter what his velocity. This property can be shown to imply that the field is zero on the average and that the fluctuations increase in magnitude at shorter wavelengths. The net result is that an observer cannot exploit the fluctuations to determine his speed.

The fluctuations can, however, serve to determine acceleration. In 1976 it was shown by William G. Unruh of the University of British Columbia that a hypothetical particle detector undergoing constant acceleration would react to the vacuum fluctuations just as if it were at rest in a gas of particles (and hence not in a vacuum) with a temperature proportional to the acceleration. An unaccelerated detector would not react to the fluctuations at all.

The idea that temperature and acceleration can be related in this way has led to a rethinking of what is meant by "vacuum" and to a recognition that there are different kinds of vacua. One of the simplest nonstandard vacua can be created by repeating, in a quantum-mechanical context, a thought experiment first proposed by Einstein. A closed elevator car is imagined to be drifting freely in empty space. A "playful spirit" starts tugging on it, bringing it into a state of constant acceleration, top end forward. The walls of the car are assumed to be perfect conductors, impermeable to electromagnetic radiation, and the car itself is assumed to be completely evacuated, so that it contains no particles. Einstein introduced this imaginary scene as a way of illustrating the equivalence of gravitation and acceleration, but a reconsideration of it shows that several strictly quantum-mechanical effects can also be expected.

To begin with, at the moment the acceleration begins the floor of the car emits an electromagnetic wave that propagates to the ceiling and then bounces back and forth. (To show why the wave is emitted would require a detailed mathematical analysis of an accelerated electrical conductor, but the effect is analogous to the creation of the acoustic compression wave that would appear if the car were full of air.) If some dissipation in the walls of the car is temporarily allowed, the electromagnetic wave is converted into photons with a thermal energy spectrum, or in other words into black-body radiation characteristic of a certain temperature.

The car now contains a dilute gas of photons. A refrigerator with a radiator on the outside can be installed to get rid of the photons, at some cost in energy from an external power supply. The end result, when all the photons have been pumped out, is a new vacuum inside the car, one that is subtly different from the standard vacuum outside. The difference consists in the following. First, an Unruh detector that shares the acceleration of the elevator car, and that would react thermally to the field fluctuations if it were placed in the standard vacuum outside, displays no reaction inside. Second, the two vacua differ in their energy content.

Specifying the energy of a vacuum requires the resolution of some delicate issues in quantum field theory. I noted above that a free field is equivalent to a collection of harmonic oscillators. The ground-state fluctuations of the oscillators give the vacuum field a residual energy known as the zero-point energy. Because the number of field oscillators per unit volume is infinite, the energy density of the vacuum would also seem to be infinite.

An infinite energy density is an embarrassment. Theorists have introduced a number of technical devices to exorcise it. The devices are part of a general program, called renormalization theory, for handling various infinities that crop up in quantum field theory. Whatever device is adopted must be universal, in the sense that it is not tailored to a spe-

Today's Chevrolet

Braggin' Wagon

Cavalier **Room for the one that didn't get away** Chevrolet Cavalier Wagon has more total room than Subaru DL, Nissan Sentra or Toyota Tercel wagons, so that means more comfort, more convenience and more cubic feet for you.

More than just more room Cavalier Wagon also gives you front-drive traction. And Cavalier's 2.0 Liter, high-compression engine with electronic fuel injection gives you more standard horsepower than the three leading import wagons.

Braggin' before it ever hits the road Before any Cavalier Wagon ever hits the road, it's already been through over 1,000 different inspections. Dedicated workers using computerized robots and lasers achieve a high level of precision fit and finish. There's even a computer to check the computer's work.

At Chevrolet we're working to bring you the cars and trucks you want and need—that's what Taking Charge is all about.

Some Chevrolets are equipped with engines produced by other GM divisions, subsidiaries, or affiliated companies worldwide. See your dealer for details.

OFFICIAL U.S. CARS AND TRUCKS OF THE XIV OLYMPIC WINTER GAMES

Sarajevo '84

© 1982 OCOG Sarajevo '84

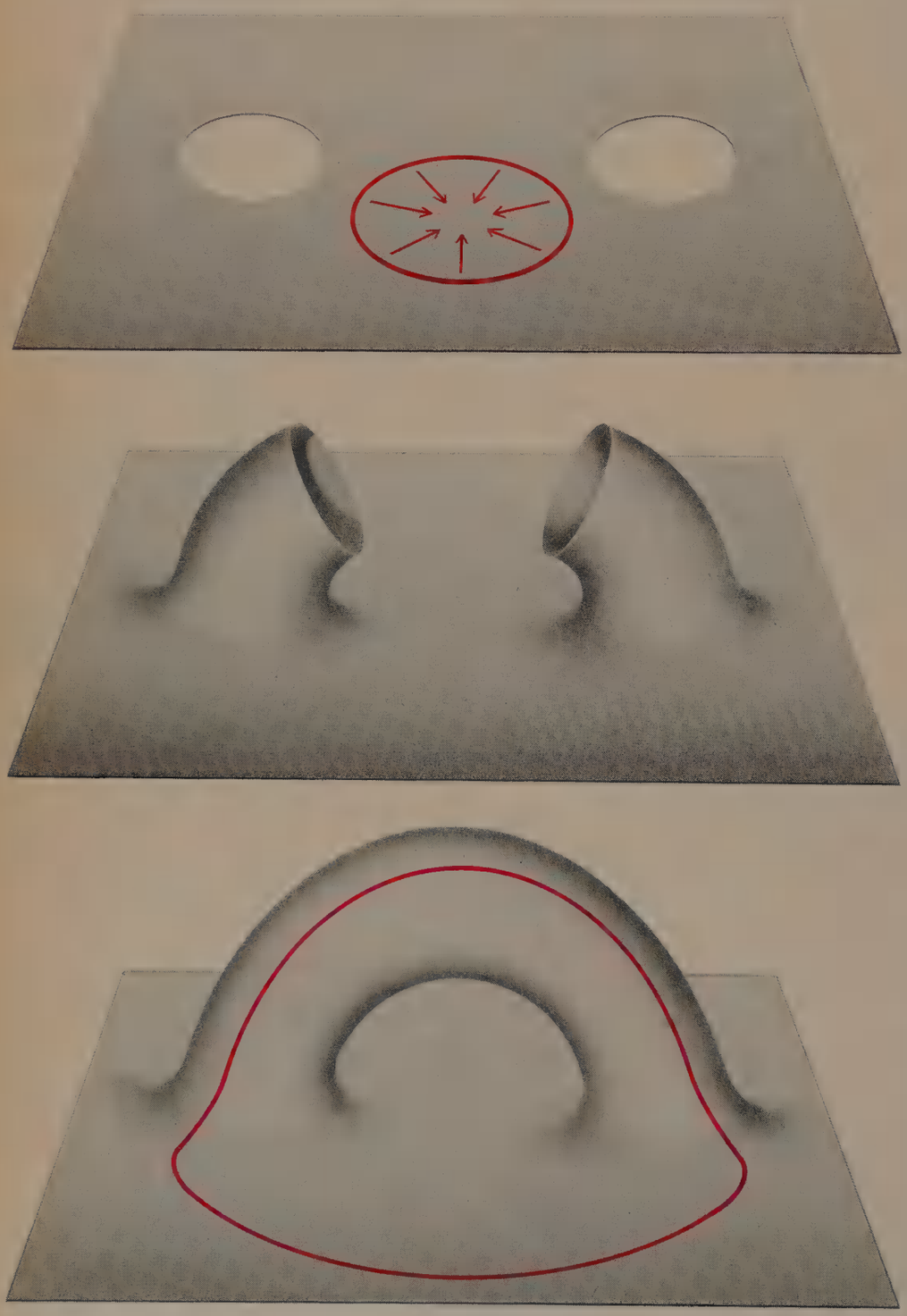




cific physical setting but can be applied uniformly in all settings. It must also yield a vanishing energy density in the standard vacuum. The latter requirement is essential for consistency with Einstein's theory because the standard vacuum is the quantum equivalent of flat, empty spacetime. If there were any energy in it, it would not be flat.

As a rule the various approaches to renormalization theory give identical results when they are applied to the same problem, which inspires confidence in their validity. When they are applied to the vacua inside and outside the elevator car, they yield zero energy density outside and a negative energy density inside. A negative vacuum energy is a surprise. What can be less than nothing? A moment's reflection, however, makes the reasonableness of the negative value apparent. Thermal photons must be added to the interior of the car to make an Unruh detector inside act as it would in the standard vacuum outside. When the photons are added, their energy brings the total interior energy up to zero, equal to that of the vacuum outside.

It must be stressed that such bizarre effects would be difficult to observe in practice. For accelerations common in daily life, even in high-speed machinery, the negative energy is far too small to be detected. There is one case, however, in which a negative vacuum energy has been observed, at least indirectly: in an effect predicted in 1948 by H. B. G. Casimir of the Philips Research Laboratories in the Netherlands. In the Casimir effect two clean, parallel, uncharged, microscopically flat metal plates are placed very close together in a vacuum. They are found to attract each other weakly with a force that can be attributed to a negative energy density in the vacuum between them.



"WORMHOLE" IN SPACETIME is a hypothetical structure that could alter the topology of the universe. On a flat plane a wormhole is formed by cutting two openings and stretching the cut edges into tubes that are then joined. On the original plane any closed curve can be shrunk to a point (*color*). A curve passing through the wormhole, however, cannot be shrunk. A wormhole in three-dimensional space or four-dimensional spacetime is not conceptually different.

The vacuum becomes even more complex when spacetime is curved. Curvature influences the spatial distribution of the quantum field fluctuations and, like acceleration, can induce a nonzero vacuum energy. Because curvature can vary from place to place, the vacuum energy can also vary, being positive in some places and negative in others.

In any consistent theory energy must be conserved. Assume for the moment that an increase in curvature causes an increase in the quantum vacuum energy. That increase must come from somewhere, and so the very existence of quantum field fluctuations implies that energy is needed to bend spacetime. It follows that spacetime resists bending. This is just like Einstein's theory.

In 1967 the Russian physicist Andrei Sakharov proposed that gravitation may be a purely quantum phenomenon, arising from vacuum energy, and suggested that Newton's constant G, or equivalently the stiffness of spacetime, may be computable from first principles. The proposal faces several difficulties. First, it requires that gravity be replaced, as a fundamental field, by some "grand unified gauge field" suggested by the known elementary particles. Here a fundamental mass must be introduced so that an absolute scale of units is still obtained; hence one fundamental constant is replaced by another.

Second, and perhaps more important, the calculated dependence of the vacuum energy on curvature turns out to yield a theory of gravity more complicated than Einstein's. Depending on the number and type of elementary fields chosen and on the method of renormalization, the vacuum energy, rather than increasing with increasing curvature, can even decrease. Such an inverse relation would mean that flat spacetime is unstable and that it would tend to wrinkle like a prune. Here I shall assume that the gravitational field is fundamental.

A true vacuum is defined as a state of thermal equilibrium at a temperature of absolute zero. In quantum gravity such a vacuum can exist only if the curvature is independent of time. When the curvature is time-dependent, particles can appear spontaneously in the vacuum (with

PROJECTION

PRESENTING THE END OF THE DARK AGES.

RCA INTRODUCES ITS BEST AND BRIGHTEST GENERATION OF BIG-SCREEN TV'S.

BIG, BRIGHT AND BEAUTIFUL. The picture you see on RCA's big 45-inch screen (measured diagonally) marks the beginning of a bright new era. Its remarkable 65% increase in peak brightness and 200% increase in contrast make it light-years ahead of earlier RCA projection sets. Watching projection television in the dark is a thing of the past.

Another improvement is a big, crowd-pleasing viewing angle that makes just about every seat in the house the best seat in the house.

VIDEO MONITOR CAPABILITY. Our high-performance giant is also a video monitor. Its convenient system of input/output jacks lets you hook up your video cassette recorder, videodisc player, and stereo for maximum enjoyment.

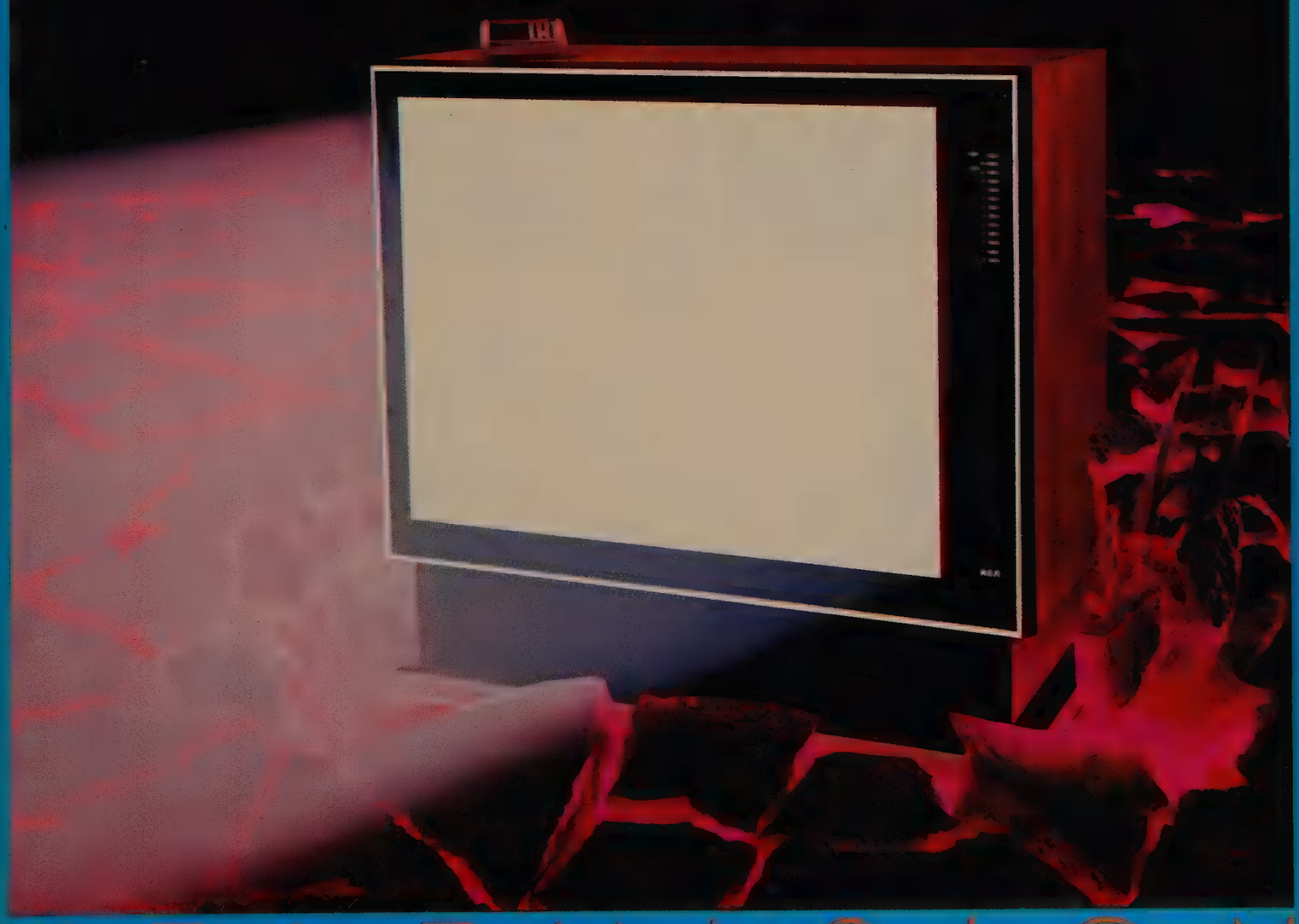
SIZE ISN'T EVERYTHING. RCA's best and brightest features our exclusive Digital Command Center. The total control remote control that enables you to switch from broadcast to VCR to videodisc viewing at the touch of a button.

All of this comes wrapped in the most compact cabinet in RCA big-screen history. To see our best generation of big-screen TV's ever, visit your RCA dealer.

It promises to be a most enlightening experience.

For more information and a free copy of "Living With Video" (a $2.50 retail value), write: RCA Consumer Electronics, Department 32-312P, P.O. Box 1976, Indianapolis, IN 46206.

WE'LL OPEN YOUR EYES RCA



TELEVISION

the result, of course, that it is no longer a vacuum).

The mechanism of particle production can again be explained in terms of harmonic oscillators. When the curvature of spacetime changes, the physical properties of the field oscillators change too. Suppose an ordinary oscillator is initially in its ground state, subject to zero-point oscillations. If one of its properties, such as its mass or the stiffness of its spring, is changed, then its zero-point oscillations must adjust themselves to the change. After the adjustment there is a finite probability that the oscillator is no longer in its ground state but in an excited state. The phenomenon is analogous to the increased vibration induced in a vibrating piano wire when its tension is increased; the effect is known as parametric excitation. In the quantum field the analogue of parametric excitation is the production of particles.

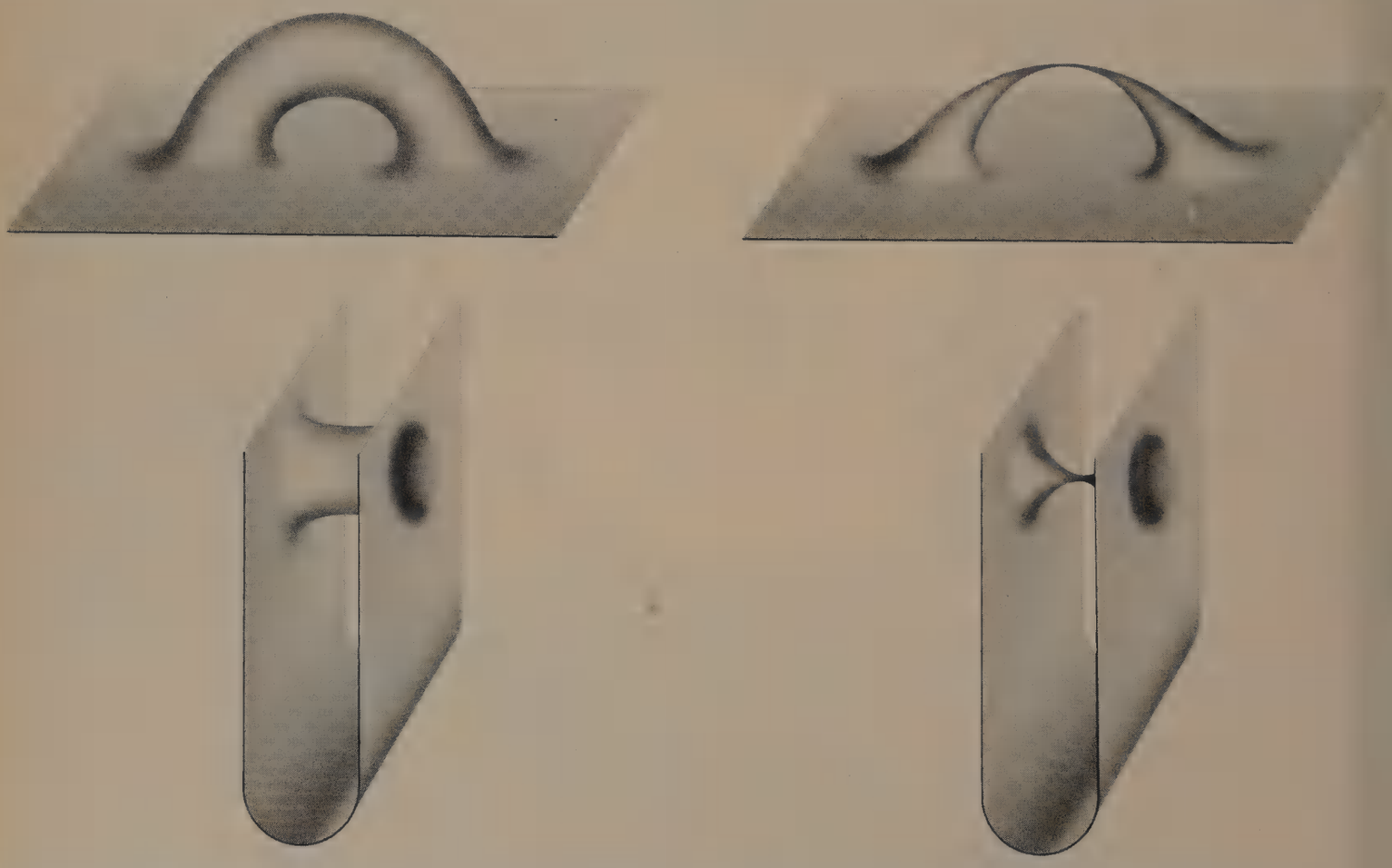
The particles generated by time-varying curvature appear randomly. It is not possible to predict in advance exactly where or when a given particle will be born. One can, however, calculate the statistical distribution of the particles' energy and momentum. Particle production is greatest where the curvature is greatest and changing most rapidly. It was probably very large in the big bang and could have had a major effect on the dynamics of the universe in its earliest moments. It is not implausible that particles created in this way could account for all the matter in the universe!

Attempts to calculate big-bang particle production were begun independently a decade ago by the Russian academician Yakov B. Zel'dovich and by Leonard E. Parker of the University of Wisconsin at Milwaukee. Many others have taken up the question since. Although several of the results are suggestive, none is definitive. Moreover, a major question hangs over the effort: What does one choose for the initial quantum state at the instant of the big bang? Here the physicist must play God. None of the proposals made so far seems uniquely compelling.

The one other event in the universe during which curvature should change rapidly is the collapse of a star to form a black hole. Here quantum-mechanical calculations have led to a real surprise, essentially independent of initial conditions. In 1974 Stephen W. Hawking of the University of Cambridge showed that the variation in curvature near a collapsing black hole creates a radiating stream of particles. The stream is steady and continues long after the black hole has become geometrically quiescent. It can continue because time seems to slow down in the enormous gravitational field near the "horizon" surface of a black hole; for an external observer all activity comes to a virtual standstill. Particles born closer to the horizon are delayed longer in their outward journey.

Although the delay in emission implies that there must be an enormous number of particles crowded near the horizon, each "waiting its turn" to escape, the total energy density in this region is actually negative and rather small. The positive energy carried by the particles is largely compensated for by an enormously negative vacuum energy that would be there if the particles were absent (for example, if the black hole had always existed and had never been created by gravitational collapse).

It can be shown that the particles emitted are statistically uncorrelated and that their energy spectrum is thermal. The black-body character of the radiation is perhaps its most important property. It allows both a temperature and an entropy to be assigned to a black hole. The entropy, which measures the



DISTANT REGIONS OF THE UNIVERSE could in principle be connected by a wormhole, suggesting that faster-than-light communications might be established between them; actually the scheme cannot work. In the drawing of a wormhole at the upper left the distance between the holes in the "outside world" is comparable to the distance through the "throat." In the wormhole at the lower left the outside distance is much greater. In the lower drawings the space represented by the plane appears to be bent, but that is only because it is viewed from the perspective of a higher-dimensional space; to an observer living in the plane it would appear approximately flat. Whether or not the throat is a short cut, it is not possible to pass through it. The reason is that a wormhole invariably connects two black holes. The throat "pinches off," as is shown at the right, and anything that enters it is crushed to infinite density before reaching the other side.

thermodynamic disorder of the system, turns out to be proportional to the area of the horizon surface. For a black hole of stellar mass it is enormous: more than 19 powers of 10 larger than the entropy of the star that collapsed to form the black hole. The temperature, on the other hand, is inversely proportional to the mass and, if the mass is stellar, is more than 11 powers of 10 smaller than that of the parent star.

Because the quantity of radiation emitted by an object depends on its temperature, the Hawking radiation from an astrophysical black hole is utterly negligible. It becomes important only for "mini" black holes, those with a mass of less than about 10^{10} grams. The only conceivable way that mini black holes could have been formed is through compression during the big bang. It is possible they were copiously produced then, in which case they would have contributed significantly to the entropy of the universe.



FLUCTUATING TOPOLOGY, which is a feature of spacetime in some attempts to formulate a theory of quantum gravity, raises serious conceptual difficulties. Shown here are two representations of a wormhole that has just pinched off, leaving two "dimples." If such an event can take place, the reverse process should also be possible; in other words, the dimples should be able to coalesce into a new wormhole. The reversal seems plausible where the dimples appear to be close together but not where they appear to be far apart. The ideas of "close" and "far," however, depend on embedding the surface in a higher-dimensional space. For an observer within the surface the objects represented by the two drawings would be indistinguishable.

The energy of particles created by a time-varying curvature cannot be conjured up from nowhere. It is taken from spacetime itself. It follows that the particles act back on spacetime. Attempts have been made to calculate this "back reaction" in the case of the big bang, to determine its dynamical effect on the early universe. One goal has been to see whether the back reaction might suppress the infinite initial density of matter required in Einstein's classical theory. The infinite density is a barrier to all further inquiry. If it could be replaced by a merely enormous density, one might ask: What was the universe doing *before* the big bang?

In the 1960's Roger Penrose of the University of Oxford and Hawking showed that Einstein's classical theory is incomplete. It predicts the past or future occurrence of infinite densities and infinite curvatures under a variety of physically reasonable present conditions. A theory that predicts an infinite value for an observable quantity ceases to be able to predict beyond that point. Since physicists believe in the ultimate comprehensibility of nature, they expect that such a theory needs to be broadened, to include a wider variety of phenomena. The conservative view at present is that the inclusion of quantum effects is the only reasonable cure in sight for the incompleteness of Einstein's theory.

Calculations of the back reaction on the big bang are done by numerical simulation with a digital computer. So far they have given ambiguous results. One difficulty has been to determine, as input for the computer, a believable value for the combined energy density of the particles produced and the quantum vacuum on which they are superposed.

The effect of the back reaction is of special importance in the case of a black hole. The Hawking radiation steals both entropy and energy from a black hole. The mass of the hole consequently decreases. The rate of decrease is slow at first but picks up as the temperature rises. Ultimately the rate of change becomes so great that the approximations made in Hawking's calculations fail. It is not known what happens thereafter. Hawking thinks that his approximations continue to be qualitatively correct and that the black hole ends its life in a spectacular flash, momentarily leaving behind a "naked singularity" in the causal structure of spacetime.

Any singularity, whether naked or not, represents a failure of the theory. If Hawking is right, not only is Einstein's theory incomplete but also the quantum theory is incomplete. The reason is that for every particle born outside the horizon surface another is born inside. The two particles are correlated in the sense that an observer could detect "probability interference effects" if he could communicate simultaneously with both particles. Hawking assumes that the particles inside get crushed to infinite density and cease to exist. The moment they cease to exist the standard probability interpretation of quantum mechanics fails. Probability itself gets lost in the infinite crush.

An alternative and equally plausible assumption is that the very framework of quantum field theory one erects around Einstein's theory prevents both probability and information from being lost in the collapse. It is quite possible the back reaction becomes so extreme that the crush is prevented from becoming infinite. The horizon, which is a mathematical construct rather than a physical one, may not be there at all as a strict one-way barrier. The matter that collapsed to form the black hole may ultimately be accounted for, particle by particle. That there will be a final flash of Hawking radiation and enormous densities inside the hole no one doubts. The very pressures to which the nuclear particles are subjected, however, may convert them into photons and other massless particles, which may finally escape, carrying with them the small remaining energy and all the quantum correlations. These end products need carry none of the original entropy of the black hole. That has all been stolen by the Hawking radiation.

I come now to the deep and difficult part of quantum gravity. When a quantum effect, such as particle production or vacuum energy, reacts back on the curvature of spacetime, the curvature itself becomes a quantum object. A consistent theoretical framework demands that the gravitational field itself be quantized. For wavelengths that are long compared with the Planck length the quantum fluctuations of the quan-

THE · 1984 · CHRYSLER



The unique LeBarons: no other series of cars offers so much.

Compare the 1984 LeBaron series to other cars that offer you luxury. No other cars—domestic or imported—combine luxury, high mileage, front wheel drive and Chrysler's famous Protection Plan* for 5 years or 50,000 miles.

Compare specifics. Every LeBaron is engineered with contemporary front-wheel drive to help give you a sense of driving confidence. Buick Regal and Oldsmobile Cutlass Supreme are rear-wheel drive. Compare further. You'll find no European or Japanese car offers the passenger room and high gas mileage of LeBaron sedan.

Insist on quality. Every LeBaron is built to Chrysler standards of excellence...and backed by Chrysler's 5-year or 50,000-mile Protection Plan on engine and

*5 years/50,000 miles, whichever comes first. Limited warranty on engine/powertrain and outer body rust through. Deductible. See dealer for details. †Use EPA est. mpg for comparison. Actual mileage may vary depending on speed, trip length

LE BARON · SERIES



They are like no other cars in America, Europe or Japan.

ower train, and protection against outer body rust-hrough for the same period.

Choose your luxury. For 1984 you can add Mark Cross eather seating ar the pirited extra performance of uel injected turbopower. And LeBaron even offers you - choice of two unique convertibles– the ultimate luxury.

The 1984 Chrysler LeBaron Series... the proud product of The New Chrysler Technology. Buckle up for safety.

Coupe opts. 41 EST HWY 27 EPA EST MPG Coupe base price $8783*†

The LeBaron Series. A product of The New Chrysler Technology.

Chrysler

y be required. Excludes leases. See dealer for details. *Late availability Sedan and Coupe. Not available on wagon. ather. Hwy. mileage probably less. CA. ests. lower. †Base sticker price. Taxes, title and destination charges extra.

tized gravitational field are small. They can be accurately represented if they are treated as a weak disturbance on a classical background. The disturbance can be analyzed the way an independent field would be. It contributes its share both to the vacuum energy and to particle production.

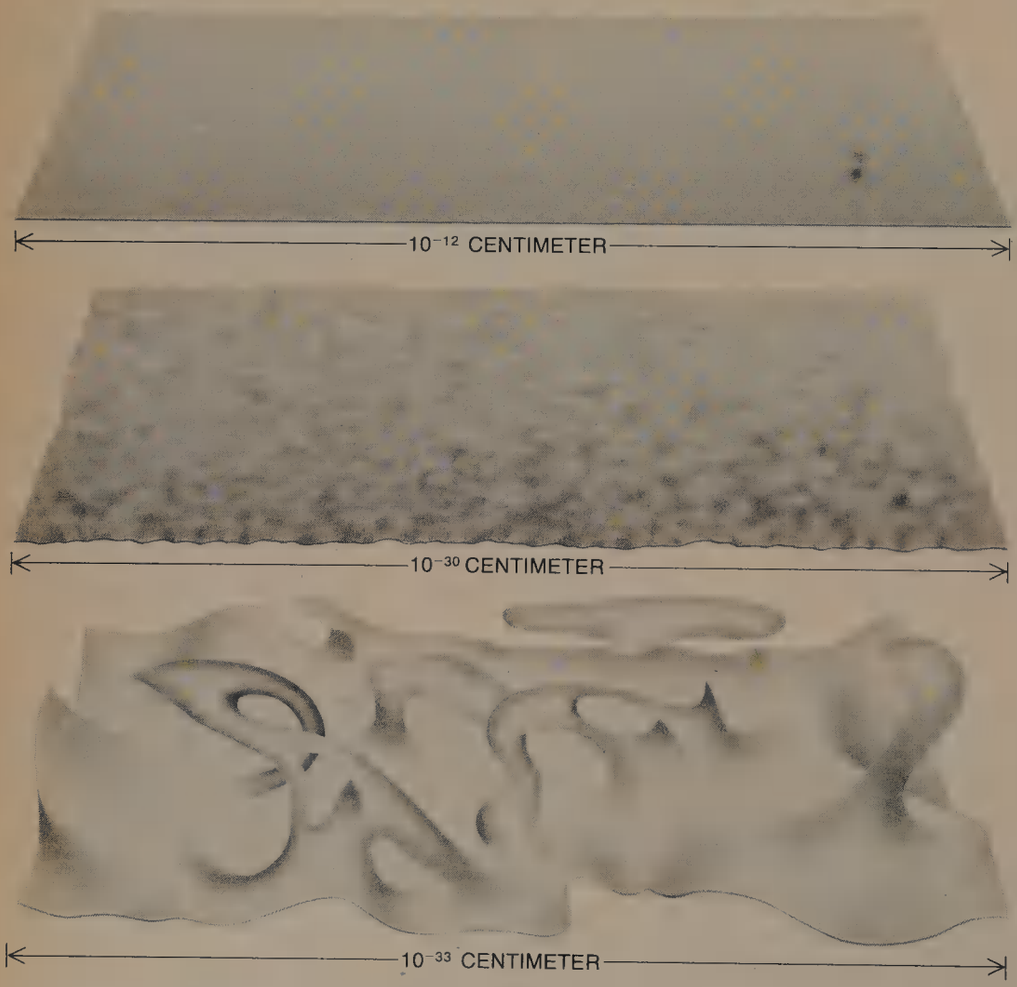
At Planckian wavelengths and energies the situation is profoundly more complicated. The particles associated with a weak gravitational field are called gravitons; they are massless and have a spin angular momentum of $2\hbar$. It is unlikely that individual gravitons will ever be observed directly. Ordinary matter, even an entire galaxy of ordinary matter, is almost totally transparent to them. Only when they reach Planckian energies do they interact appreciably with matter. At these energies, however, they are capable of inducing Planckian curvatures in the background geometry. The field with which they are associated is then no longer weak and the very concept of "particle" is ill-defined.

At long wavelengths the energy carried by a graviton distorts the background geometry. At shorter wavelengths it distorts the waves associated with the graviton itself. This is a consequence of the nonlinearity of Einstein's theory: when two gravitational fields are superposed, the resulting field is not equal to the sum of the two components. All nontrivial field theories are nonlinear. For some it is possible to deal with the nonlinearity by a method of successive approximations called perturbation theory, a name taken originally from celestial mechanics. The essence of the method is to refine an initial approximation by making a series of progressively smaller corrections. When perturbation theory is applied to quantized fields, it leads to infinities that must be eliminated by renormalization.

In the case of quantum gravity perturbation theory does not work. There are two reasons. First, at Planckian energies the successive terms of the perturbation series (that is, the successive corrections) are comparable in magnitude. Truncating the series at a finite number of terms does not yield a valid approximation; instead the entire infinite series must be summed. Second, the individual terms of the series cannot be consistently renormalized. In each order of approximation new classes of infinities appear, which have no counterparts in ordinary quantum field theory. They arise because in quantizing the gravitational field one quantizes spacetime itself. In ordinary quantum field theory spacetime is a fixed background. In quantum gravity the background not only reacts to the quantum fluctuations but also shares in them.

In a narrow technical response to these difficulties a few attempts have been made to sum infinite subsets of terms of the perturbation series. The results, notably the complete disappearance of infinities, are both suggestive and encouraging. The results must be viewed with caution, however, because severe approximations are made in obtaining them and the perturbation series is never summed in toto. They are nonetheless being used to compute improved estimates of the effect of the back reaction on the big bang.

From a broader point of view one must expect other problems to arise whose resolution cannot even be approached by summing series. A quantized spacetime is one whose causal structure is fluctuating and uncertain. At Planckian dimensions the very distinction between past and future becomes blurred. By analogy with the phenomenon of tunneling in atomic systems, which allows an electron to pass through an energy barrier it cannot climb over, one must expect processes that are not allowed in Einstein's classical theory, including faster-than-light travel over Planckian distances. How the probabilities of such processes should be calculated is largely unknown. In many cases it is not even known which are the right questions to ask. There are no experiments to guide us. It is therefore still possible to indulge in flights of fancy.



QUANTUM VACUUM, as envisioned by John Archibald Wheeler in 1957, becomes increasingly chaotic as one inspects smaller regions of space. At the scale of the atomic nucleus (*top*) space looks very smooth. At dimensions of 10^{-30} centimeter (*middle*) a certain roughness in the geometry begins to appear. At the scale of the Planck length, 1,000 times smaller still (*bottom*), the curvature and the topology of space are continually undergoing violent fluctuations.

One of the most persistent flights of fancy, referred to repeatedly in the literature on quantum gravity, is the idea of fluctuating topology. The basic notion, which was introduced by Wheeler in 1957, is the following. The vacuum fluctuations of the gravitational field, like those of all other fields, increase in magnitude at shorter wavelengths. If standard weak-field results are extrapolated into the Planck domain, the curvature fluctuations become so violent that they appear to be capable of tearing holes in spacetime and changing its topology. Wheeler imagines the vacuum as being in a state of perpetual turmoil, with wormholes (and more complicated structures) of Planckian size continually forming and disappearing. The turmoil is "visible" only at the Planck level. At a coarser level spacetime continues to appear smooth.

An immediate objection can be raised: Every topological change is necessarily accompanied by a singularity in the causal structure of spacetime, so that one is faced with the same difficulty that arises in Hawking's view of black-hole decay. Suppose, however, Whee-

ler's view is correct. One of the first questions that must be asked is: How much do the topological fluctuations contribute to the energy of the vacuum and how do they affect the resistance of spacetime (at the coarse level) to bending? To date no one has given a convincing answer, largely because no coherent picture has emerged of the topological transition process itself.

To appreciate just one of the obstacles in the way of constructing such a picture, consider the process shown in the illustration on page 123. The drawing gives two representations of the same event: a wormhole has just pinched off, leaving two residual "pseudopods" on a simply connected space. In one representation space is shown bent, in the other it is shown flat.

Now consider the reverse process: the formation of a wormhole. If there is a finite probability for a wormhole to disappear by pinching off, there is also a finite probability for one to form. Here a new difficulty arises. From the time-reversed point of view the illustration represents two pseudopods that have grown spontaneously in the quantum vacuum. In one representation the possibility that the two pseudopods may join to form a wormhole seems reasonable. In the other it does not. And yet the physical situation is the same in both drawings. The formation of the wormhole seems reasonable in the one case because the pseudopods appear to be close together. The "closeness," however, is not an intrinsic property of the spatial arrangement, as the other representation clearly shows. A notion of "closeness" requires the existence of a higher-dimensional space in which spacetime is embedded. Furthermore, the higher-dimensional space must be endowed with physical properties in order for the pseudopods to transmit a sense of nearness to each other. But then spacetime is no longer the universe. The universe is something more. If one remains faithful to the view that the properties of spacetime are intrinsic and are not the result of something outside, a coherent picture of topological transitions seems beyond reach.

Another difficulty with topological fluctuations is that they might undermine the macroscopic dimensionality of space. If wormholes can form spontaneously, the wormholes themselves can develop wormholes, ad infinitum. Space could evolve into a structure that, although three-dimensional at the Planck level, has four apparent dimensions or more at a larger scale. A familiar example of this process is the formation of a foam, which is made up entirely of two-dimensional surfaces but has a three-dimensional structure [*see illustration on this page*].

Because of difficulties such as these some physicists have suggested that the conventional description of spacetime as a smooth continuum fails at the Planck level and must be replaced by something else. What that something else consists of has never been made very clear. In view of the success of the continuum description over scales of length spanning more than 40 powers of 10 (60 if the possible failure is assumed to come only at the Planck level), it would seem equally reasonable to suppose that the continuum description is valid at all levels and that topological transitions simply do not exist.



DIMENSIONALITY OF SPACE is put into question by the possibility that spacetime has a complex topology. The surface shown is two-dimensional, but its topological connections give it the appearance of a three-dimensional object. It is conceivable that three-dimensional space perceived at a macroscopic scale actually has fewer dimensions but is topologically convoluted.

Even if the topology of space is unchanging, it is not necessarily simple, even at a microscopic level. It is conceivable that space could have a foam structure, built in from the beginning, in which case its apparent dimensionality could be greater than its real dimensionality. Its apparent dimensionality could also be less than its real dimensionality.

The latter possibility was proposed in a theory put forward by Theodor Kaluza in 1921 and by Oskar Klein in 1926. In the Kaluza-Klein theory space is four-dimensional and spacetime is five-dimensional. The reason space *appears* to be three-dimensional is that one of its dimensions is cylindrical, as in the universe discussed above, but with an important difference: the circumference of the universe in the cylindrical direction, instead of being billions of light-years, is only a few (perhaps 10 or 100) Planck units. As a result an observer who attempts to penetrate the fourth spatial dimension is almost instantly back where he started. Indeed, it is meaningless to speak of such an attempt, because the very atoms of which the observer is composed are vastly larger than the cylindrical circumference. The fourth dimension is simply unobservable as such.

It can nonetheless manifest itself in another way: as light! Kaluza and Klein showed that if their five-dimensional

Over a million people use Microsoft® software on their microcomputers. There's a reason.

Powerful simplicity. That's the concept behind every microcomputer software program we write. Powerful programs that allow you to do more with less effort. Programs that let you spend more time thinking about your problem...and less time thinking about the computer. It's a good reason why over a million people depend on Microsoft software in their everyday computing... and why the number is growing. Rapidly.

The value of experience. Microsoft made a giant step toward putting the power of computers in the hands of non-technical people when we wrote the first BASIC programming language for the very first microcomputer. It's an English-like programming language simple enough to teach in grade schools. Yet, it's so powerful that more application programs have been written in Microsoft BASIC than any other programming language.

Tools that help you think. Microsoft is an old-timer in one of the world's youngest industries. Microsoft BASIC was just the beginning. Today, we offer a broad range of proven tools for microcomputers. A complete line of programming languages. From the original BASIC to Pascal. Operating systems like XENIX™ and MS™-DOS that allow your computer to utilize a wide variety of programs. Products like the Microsoft SoftCard™ system which dramatically extends the capabilities of Apple® II computers. Management software such as Multiplan,™ which allows you to explore more alternatives to reach more intelligent business decisions. Better tools. Because they're powerfully simple.

Building a library. The great majority of microcomputers come equipped with Microsoft software. But you'll undoubtedly want more. Specialized programs that solve your particular problems. Programs for your industry, profession, or home. Programs that do more and ask less. Look to Microsoft, or to programs written in Microsoft languages. Software that's proven on millions of computers. Worldwide.

BETTER TOOLS FOR MICROCOMPUTERS

MICROSOFT®

MICROSOFT CORPORATION
10700 NORTHUP WAY
BELLEVUE, WASHINGTON 98004

Microsoft is a registered trademark of Microsoft Corporation
Multiplan, SoftCard, XENIX and MS are trademarks of Microsoft Corporation
Apple is a registered trademark of Apple Computer, Inc.



spacetime is treated mathematically in exactly the same way that four-dimensional spacetime is treated by Einstein, their theory is equivalent to Maxwell's theory of electromagnetism combined with Einstein's theory of gravitation. The components of the electromagnetic field are implicit in the equation satisfied by the curvature of spacetime. Kaluza and Klein thus invented the first successful unified field theory, a theory that gave a geometric explanation of electromagnetic radiation.

In a sense the Kaluza-Klein theory was too successful. Although it united Maxwell's and Einstein's theories, it predicted nothing new and so could not be tested against other theories. The reason was that Kaluza and Klein put constraints on the way spacetime is allowed to curve in the extra dimension. If the constraints had been removed, the theory would have predicted new effects, but the effects did not seem to correspond to reality. Hence the theory was viewed for many years as a beautiful curiosity and relegated to a closet.

The Kaluza-Klein theory was brought out of the closet in the 1960's when it was realized that the new gauge theories that were attracting increasing attention could be reformulated as Kaluza-Klein theories in which space is endowed with not merely one but several extra microscopic dimensions. It began to appear that all of physics might be explainable in geometric terms. It then became important to ask what happens if the constraints on the curvature in the compact dimensions are removed.

One thing that happens is that curvature fluctuations are predicted in the extra dimensions; the fluctuations are manifested as massive particles. If the circumference of the extra dimensions is 10 Planck units, the mass of the associated particles is roughly a tenth of the Planck mass, or about a microgram. Because the energy needed to create such particles is enormous, they almost never get produced. It therefore makes little practical difference whether the constraints on curvature fluctuations are imposed or not. Problems do remain. The chief one is that the extreme curvature of the extra dimensions gives rise to a very large energy density in the classical vacuum. The large vacuum energy is ruled out by observation.

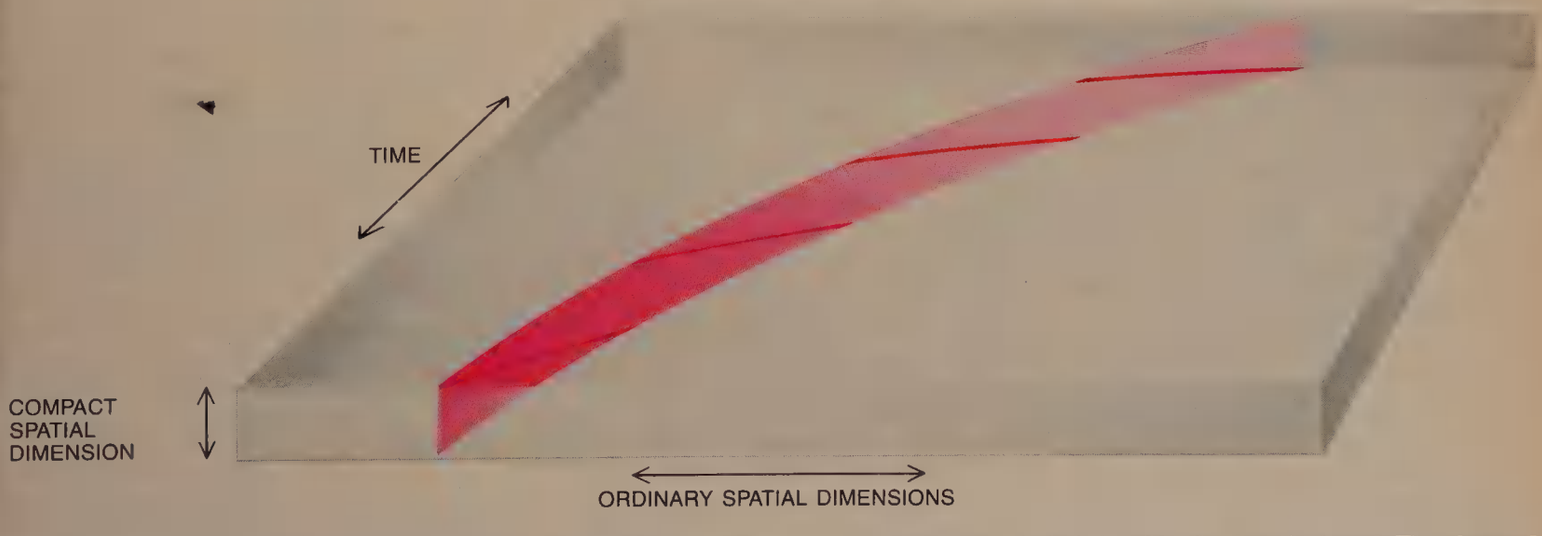
Kaluza-Klein models have never been pursued with undivided attention, and their role in physics is still uncertain. In the past two or three years, however, they have again been much scrutinized, this time in connection with the remarkable generalization of Einstein's theory known as supergravity, invented in 1976 by Daniel Z. Freedman, Peter van Nieuwenhuizen and Sergio Ferrara and (in an improved version) by Stanley Deser and Bruno Zumino.

One of the inadequacies of standard Kaluza-Klein models is that they predict the existence only of particles with spin angular momenta of 0, $\hbar$ and $2\hbar$, and even these particles are either massless or supermassive. Nowhere is there room for the particles of ordinary matter, most of which have spin angular momentum $\frac{1}{2}\hbar$. It turns out that if Einstein's theory is replaced by supergravity and if spacetime is given the Kaluza-Klein treatment, a true union of all the spin varieties is achieved.

In the "super" Kaluza-Klein model that is currently most popular seven extra dimensions are added to spacetime. These dimensions have the topology of a 7-sphere, a space that has some fascinating properties of its own. The resulting theory is extraordinarily rich and complicated, specifying huge supermultiplets of particles. The masses of the particles are still either zero or extremely large, but it is possible that a "breaking" of the symmetry of the 7-sphere will give some of the particles a more realistic mass. The large energy of the classical vacuum also persists, but it might be canceled by a negative quantum vacuum energy. It is not yet known whether these strategies for rectifying the theory will succeed. Indeed, much work will be needed to determine exactly what the theory actually states.

If Einstein could come back in spirit to witness what has become of his theory, he would certainly be astonished, and I think pleased. He would be pleased that physicists at last, after years of hesitation, have come to accept his view that theories that are mathematically elegant deserve to be studied even if they do not seem to correspond immediately to reality. He would also be pleased that physicists now dare to hope a unified field theory may be attainable. He would be particularly pleased to find his old dream that all of physics may be explainable in geometric terms seems to be coming true.

Above all he would be astonished. Astonished that the quantum theory still stands pristine and unmoved in the midst of it all, enriching field theory and itself being enriched by it. Einstein never believed the quantum theory expresses ultimate truth. He never reconciled himself to the indeterminism it implies and thought it would someday be replaced by a nonlinear field theory. Exactly the opposite has happened. The quantum theory has invaded Einstein's theory and transmuted it.



ADDITIONAL SPATIAL DIMENSIONS, beyond the three known ones, could exist if they had a "compact" form. For example, a fourth spatial dimension could be rolled up into a cylinder with a circumference of perhaps 10^{-32} centimeter. Here the hypothetical compact dimension has been "unrolled" and is represented as the vertical axis in a spacetime diagram. The path of a particle can therefore have a cyclical component; each time it reaches the maximum extent of the compact dimension it finds itself back where it started. The observed path is a projection of the actual path onto the macroscopic spacetime dimensions. If the path is a geodesic one, it can have the appearance of an electrically charged particle moving in an electric field. A theory of this kind was introduced in the 1920's by Theodor Kaluza and Oskar Klein, who showed it could explain both gravitation and electromagnetism. There has been a resurgence of interest in such theories.

Prey Switching in a Simple Ecosystem

On the island of Newfoundland, which has few species of mammals, lynxes preyed on snowshoe hares until the hare population "crashed." Then the lynxes switched to caribou calves, and the cycle continues

by Arthur T. Bergerud

In most parts of the world the ecological chain along which matter and energy are transferred from simple photosynthetic organisms to carnivorous mammals by way of the intervening herbivores is an intricate one. In any small area there are generally many species of prey and predators. For each predator there are several types of prey and for each prey several types of predators. In some regions, however, geographic conditions have led to the formation of a relatively simple ecological system including only a few species. Such ecosystems can provide a laboratory for the study of the principles that govern the size of animal populations. One of the ecosystems is on the large island of Newfoundland, where there are only 14 indigenous mammals.

The array of animals in Newfoundland is not only small but also unbalanced: nine of the 14 species are carnivorous. In a region where there are so many carnivores and so few herbivores for them to eat, any shift in the prey species can have strong reverberations; the damping of the effects that would take place in a more complex ecosystem is absent. Such a dramatic effect was observed in Newfoundland early in this century when the caribou and the arctic hare, two of the chief prey animals on the island, began to decrease in number. By means of a long and thorough investigation it was determined that in both instances the cause of the decrease was a rapid increase in the number of lynxes.

For many centuries the lynx was rare in Newfoundland. In the 1860's, however, the snowshoe hare was introduced to the island, and with this species as a source of food the lynx population quickly expanded. By the early part of this century the snowshoe hares had reached their ecological limits. The snowshoe hare population "crashed," that is, diminished rapidly, leaving the lynx without its main food item. As a result the lynx, a resourceful predator, began to prey on caribou calves and the other species of hare: the arctic hare. When the snowshoe hare increased again, the lynx returned to preying on it.

The cycle has been repeated many times. Thus in recent years the lynx, the snowshoe hare and the caribou have been connected in a triangle in which the populations of the two prey species oscillate in alternating cycles of increase and decrease. The lynx also limits the number of arctic hares, once numerous in Newfoundland but now rare. The introduction of a single new species into the simple ecosystem of the island has had a long, tangled skein of consequences for the equilibrium of the system. The unraveling of this skein could illuminate the dynamics of animal populations in finer-spun ecological webs.

The ecological system of Newfoundland is quite distinct from that of the nearby parts of the Canadian mainland. Several factors have contributed to this isolation. There is evidence that when the Wisconsin ice sheet advanced some 18,000 years ago, the ice did not cover all Newfoundland. Part of the island could have been what is called a refugium: an area where local mammals could survive. As a result some native species were preserved and continued to evolve independently of the mammals on the mainland. Nine of the 14 indigenous mammals in Newfoundland are distinct enough from those on the mainland to be classified as subspecies, which is evidence of thousands of years of genetic isolation.

Some species, particularly predators, could have reached the island on ice floes or ice bridges in the postglacial period. The sea, however, has continued to act as a barrier to a full biological exchange between the island and the coastal areas of mainland Canada. In Labrador, which is only 17.6 kilometers from Newfoundland across the Strait of Belle Isle, there are 34 mammalian species. On Cape Breton Island, which is 112 kilometers southwest of Newfoundland and is connected to the mainland, there are 38 species.

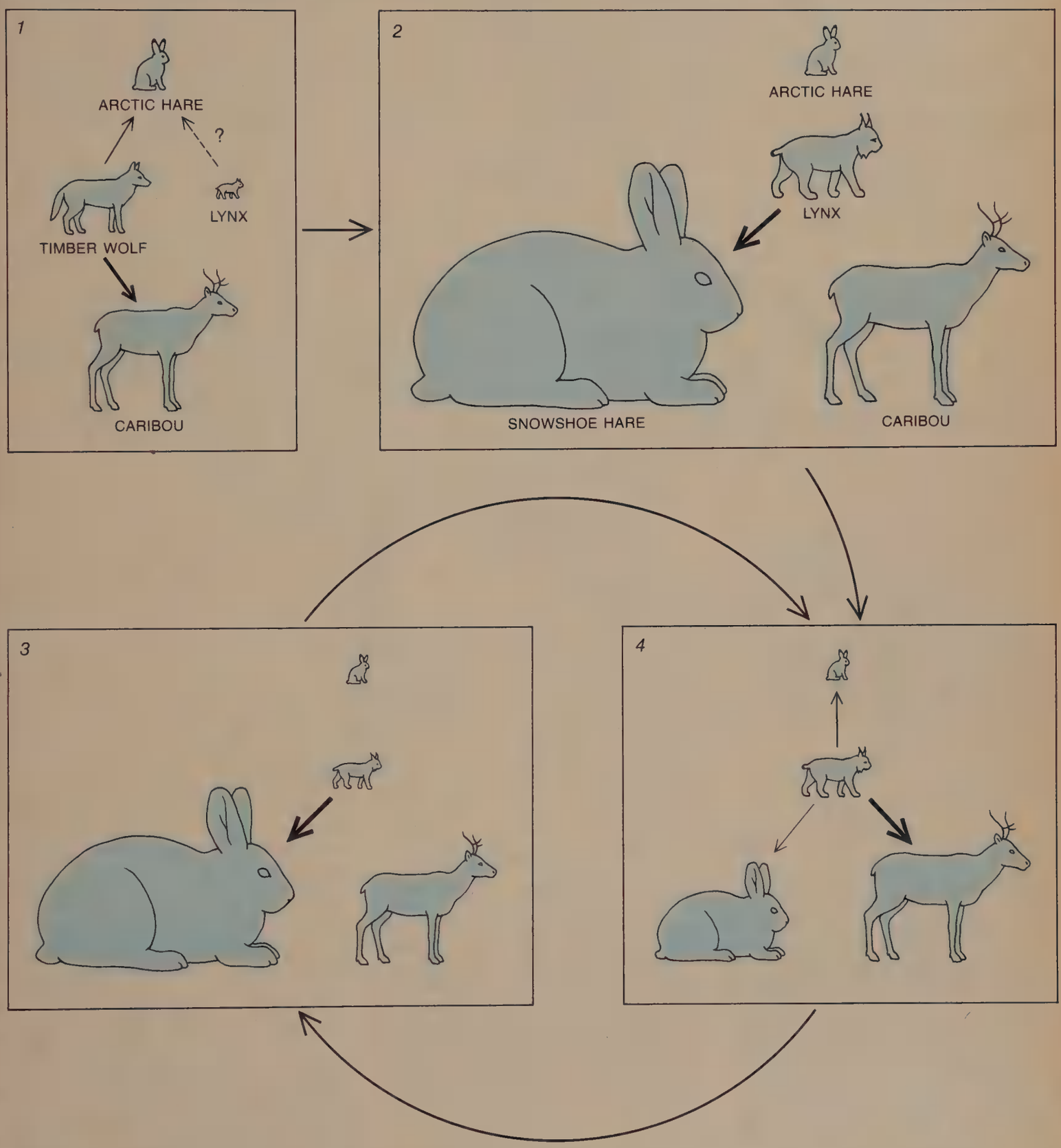
The sea around Newfoundland has acted not only as a barrier, reducing the overall number of species, but also as a filter, selecting certain types of mammals and preventing others from reaching the island. The adventurous predators were more likely than their prey to colonize the island. There were originally seven carnivorous quadrupeds on the island: the timber wolf (*Canis lupus*), the red fox (*Vulpes vulpes*), the lynx (*Lynx lynx*), the otter (*Lutra canadensis*), the black bear (*Ursus americanus*), the weasel (*Mustela erminea*) and the pine marten (*Martes americana*). The five herbivores on which these predators fed before the arrival of the snowshoe hare are the Canadian beaver (*Castor canadensis*), the North American meadow mouse (*Microtus pennsylvanicus*), the caribou (*Rangifer tarandus*), the arctic hare (*Lepus arcticus*) and the muskrat (*Ondatra zibethica*). There are also two species of bats.

In this array the usual pyramid of species is inverted. Since biological systems, like all other systems of energy transfer, have a limited efficiency, the total mass of predators is much less than the total mass of prey. By the same token the number of predator species is generally smaller than the number of prey species. Simple inverted faunas such as the one in Newfoundland are fragile, partly because they constitute an exception to the rules that normally govern food chains. My own experience in Newfoundland shows just how vulnerable such a system can be. From 1956 to 1967 I served as provincial biologist for the Newfoundland Division of Wildlife. My main task was to find the cause of a sharp decrease in the number of caribou. What I found was also relevant to other species.

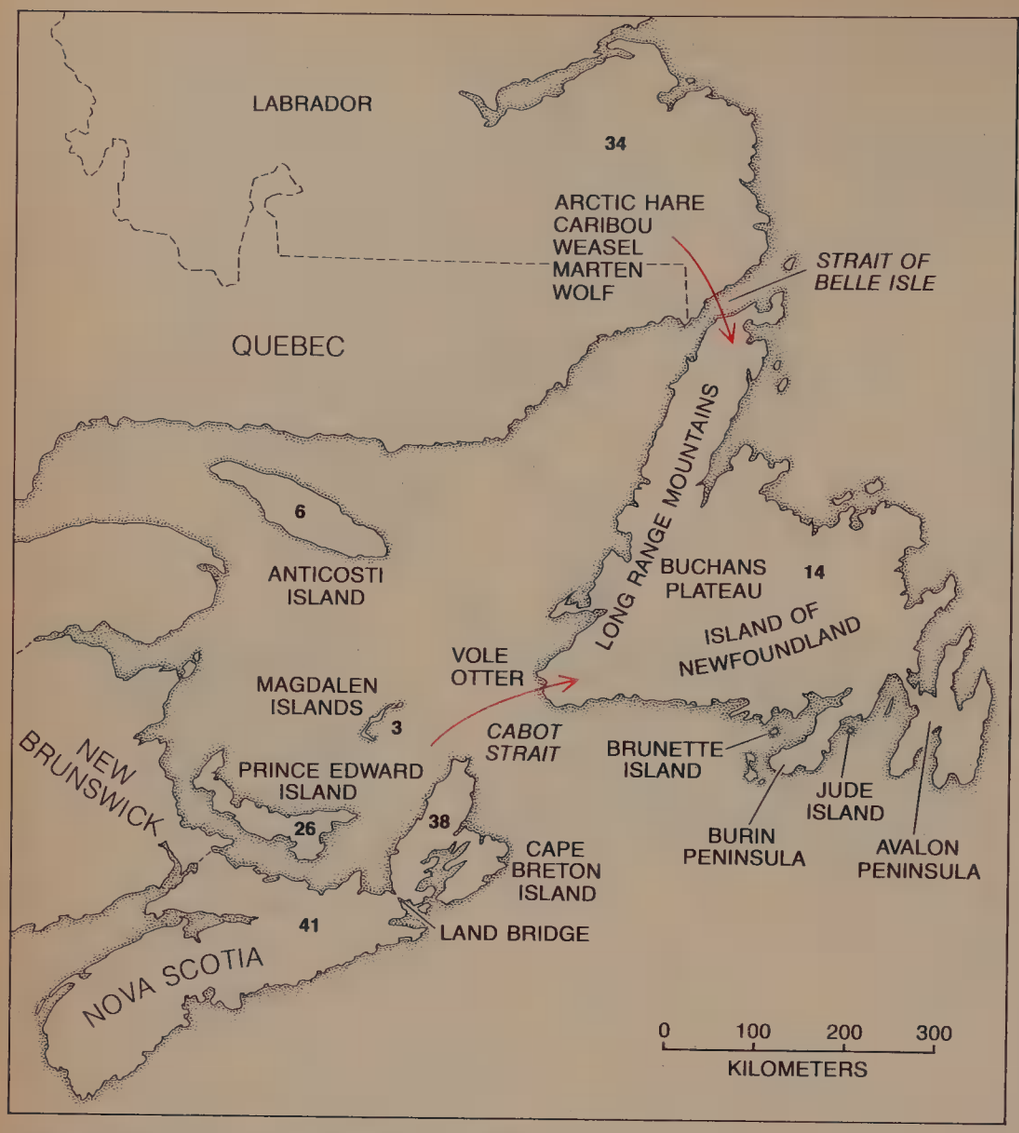
In 1900, when the interior of New-

foundland was uninhabited except for the small tribe of Micmac Indians, the caribou herds were quite large. A. A. Radclyffe Dugmore and J. G. Millais, two early naturalists, estimated that at the end of the 19th century the caribou on the island numbered between 150,000 and 200,000. These estimates, however, could be no more than educated guesses; Dugmore and Millais had made no attempt to count or sample the caribou herds systematically. On the basis of interviews with older hunters, records of the caribou harvest and the size of the available range, I have estimated that the population of caribou in Newfoundland in 1900 was 40,000, still a considerable number of animals for an island with an area of only 100,000 square kilometers.

Only 25 years later the caribou was



PREY-SWITCHING CYCLE in Newfoundland is shown schematically. For millenniums up to the middle of the 19th century the main predator on the caribou was the timber wolf. There were few native lynxes (*1*). Snowshoe hares were brought to the island as a source of food in 1864. They multiplied rapidly, reaching a peak in about 1900 (*2*). The lynxes, feeding on snowshoe hares, also multiplied rapidly. The wolf became extinct in about 1911. In about 1915 the snowshoe hare population crashed and the lynxes switched to caribou calves and arctic hares, considerably reducing the number of both prey species (*3*). Meanwhile the snowshoe hares had begun an intrinsic 10-year cycle of population growth and decline. When the hare population peaked, the lynx switched back to them, allowing the caribou population to grow again (*4*). For decades the lynx has been switching between the caribou and the snowshoe hare in a 10-year cycle. Arctic hares have become quite rare; they now live mostly in upland areas and their population fluctuates less than the other populations.



ORIGINAL FAUNA of Newfoundland includes only 14 species of mammals. Nearby regions of mainland Canada have many more species. Five species in Newfoundland may have come from Labrador on ice floes or ice bridges; two may have come from the Maritime Provinces. The sea has been a barrier to species and a filter of the types of animals that reached the island. As a result the food-chain pyramid of species in Newfoundland is inverted: nine of the 14 species are carnivorous and only five are herbivorous. Such an inverted fauna is fragile.

almost extinct. In about 1925 Dugmore estimated that there were only 200 animals left. This remarkable decrease took place even though the timber wolf, which for millenniums had been the chief predator on the caribou, became extinct in about 1911.

In the late 19th century and the early 20th the native arctic hare population also decreased from its original size, although the original number of arctic hares and the rate of decrease are not as well documented as those of the caribou. The number of arctic hares that lived in Newfoundland before the island was heavily settled can only be guessed at, but most observers agree that the territory in which arctic hares lived once extended down the Northern Peninsula to Buchans Plateau and the southern Long Range Mountains. From there the range stretched east along the south coast to the Burin and Avalon peninsulas. This domain includes much of the area of the island.

The clearest sign of a major drop in the population of arctic hares is the reduction in the hare's range since 1900. Outram Bang, a naturalist who described seven of Newfoundland's 14 native species, wrote in 1913: "The arctic hare is now very rare and local in Newfoundland, occurring only on the tops of the high mountains." By the 1950's the arctic hares were limited to the highlands of the northern and southern Long Range Mountains and Buchans Plateau and extended along the south coast to the Bay d'Espoir. Even in this area the animals were rare. The total arctic hare population of Newfoundland in recent decades has probably been less than 1,000.

When I arrived on the island, there was much more evidence on the decrease of the caribou than there was on that of the arctic hares. The investigation of the caribou's decline concentrated initially on the factors that influence the mortality rate among caribou calves. In many species of mammal the young age classes are the weakest link in the survival of the population. Indeed, the reduction in the caribou herds appeared to be due in large part to an increased death rate among the calves.

The adult males in the caribou herd spend much of the year foraging apart from the females. In mid-October the sexes meet for a mating period that lasts for about 10 days. In this time some 80 percent of the cows are impregnated. Among the females that are not impregnated some become pregnant during a shorter period of sexual receptivity about 10 days later.

The gestation period of the caribou is approximately 229 days. Because of the synchronization of mating, almost all the calves are born in a 14-day period that begins around May 24. The females calve in large groups at specific sites, and they generally return to the same calving ground every year. Most of the calving grounds in Newfoundland are open, marshy areas surrounded by scrub forest. When the calving season ends, some herds scatter into the woods to escape flying insects.

Each spring from 1957 through 1964 biologists found many dead calves on the calving grounds. The remains were easily found because the mother stayed near the calf for several days after it had died. Many of the dead calves had abscesses on their neck resulting from infection by a common pathogenic bacterium: *Pasteurella multocida* mucoid type A. The immediate cause of death was septicemia, or blood poisoning, from the infection.

Since the female caribou calve on a well-synchronized annual cycle and the fraction of females that give birth is fairly constant from year to year, it is not hard to calculate the number of new members that would be added to a caribou herd each year in the absence of perturbing influences. On the basis of such calculations it appeared that the mortality rate among caribou calves was quite high and that the calves whose remains were found represented only a minority of those that had died. From 1958 through 1967 an average of 27 percent of the calves disappeared within two weeks of their birth. By counting females with calves and those without calves it was estimated that by October about 70 percent of the females that had given birth were missing their offspring. After October, however, the death rate dropped sharply; most of the calves that made it to October survived the winter. It was the mortality in the first four months of life that was limiting the size of the caribou herds.

In addition to the missing calves and the concentration of the deaths in the spring there were other puzzles. First, at birth the ratio of male calves to female calves was 52 to 48. By the fall the ratio was reversed and there were 62 females

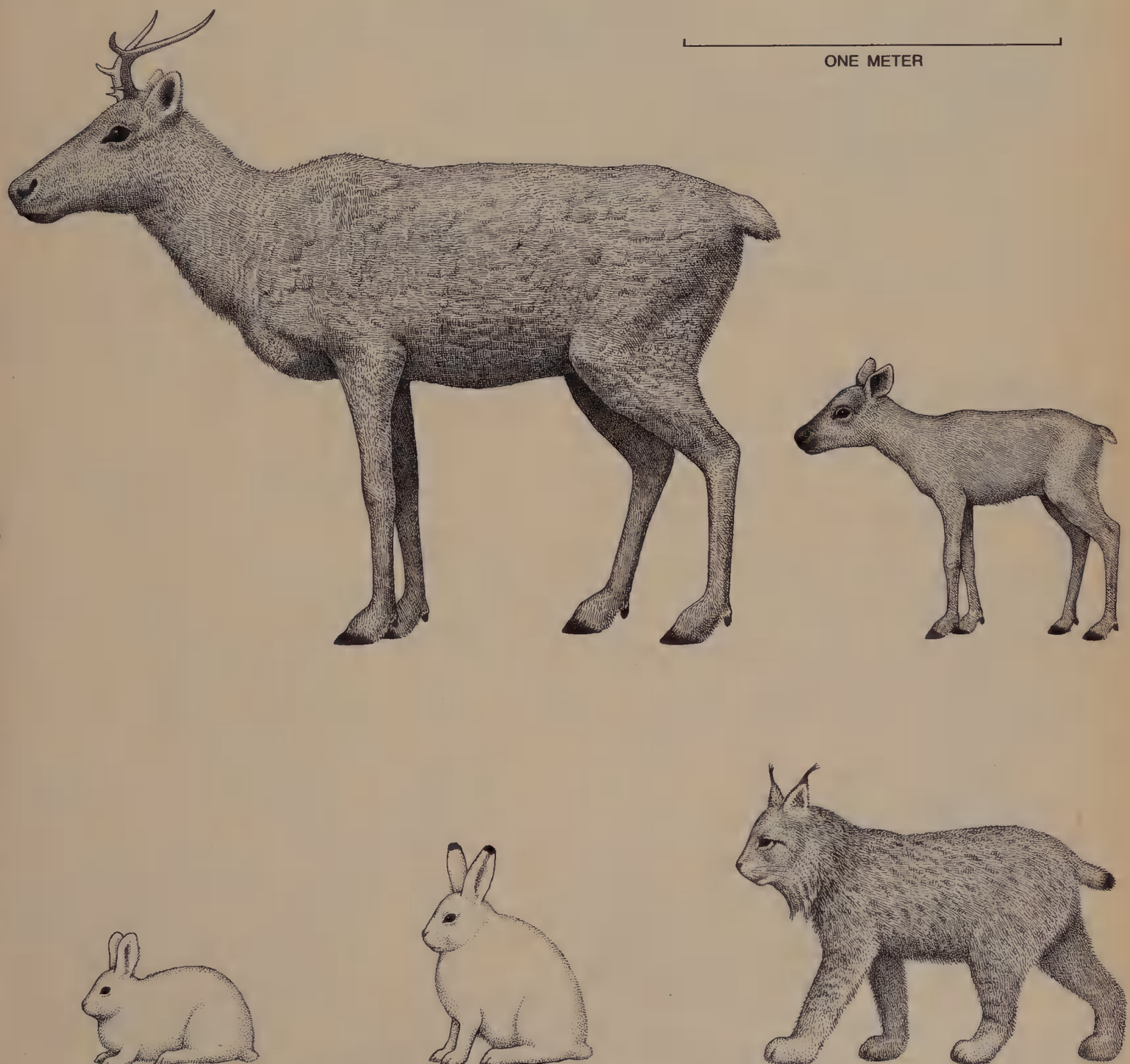
to 38 males. Second, the overall mortality was greatest in the calving grounds at low elevations. This was an unexpected finding because in such areas the wind-chill factor was at its lowest. In the alpine areas, where the climate was considerably harsher, the death rate was lower. Most intriguing of all, as the herds were counted a cyclical pattern of survival began to emerge. Annual counts showed that the fraction of calves surviving increased in 1958 and 1959, then declined steadily from 1960 through 1963.

The pieces of the puzzles began to fall into place in 1964 when the cause of the abscesses was discovered. In the spring of that year a dead calf was found with four puncture wounds in its neck but no abscesses. It seemed probable that the other dead calves had also had puncture wounds but that the wounds had been concealed by the abscesses.

The limited fauna of Newfoundland made the search for the cause of the punctures relatively simple. Among the local predators only the canine teeth of the lynx were of the right size and arrangement to cause the four evenly spaced wounds. When the skull of a lynx was obtained for verification, the lynx canines fit neatly into the punctures.

Saliva from the mouth of a lynx was cultured and was found to contain *Pasteurella multocida*. Next three caribou calves were put in a pen with a lynx. The lynx seized each calf in turn by the neck. Observers kept the predator from killing its prey, and the calves were removed from the pen. Abscesses with *Pasteurella* infections developed quickly, and the calves died respectively four, five and 15 days after being bitten.

This piece of detective work suggest-



FOUR SPECIES IN THE PREY-SWITCHING CYCLE of Newfoundland are shown. The caribou (*Rangifer tarandus*) has a synchronized cycle of births. Calves are born on long-established calving grounds during a two-week period in May and June. In the 1950's and 1960's many calves were found dead on the calving grounds with abscesses on their neck. The primary defenses of the artic hare *Lepus arcticus* (*bottom middle*) and the snowshoe hare *Lepus americanus* (*bottom left*) are alertness and speed. The snowshoe hare inhabits the boreal forest, which in winter has loose-packed snow. The arctic hare lives on the windswept tundra, which has hard-packed snow. Both hares are shown in their winter coloration. The lynx (*Lynx lynx*) stalks its prey from the underbrush and seizes the prey by the neck.



CAUSE OF ABSCESSES on the dead caribou calves was discovered in 1964. The body of a calf with four puncture wounds in the neck but no abscesses was found. When a lynx skull was obtained, the four large canine teeth fit neatly into the wounds. Culturing saliva from a lynx yielded the pathogenic bacteria that had previously been found in the abscesses. It was evident that predation by lynxes was responsible for the mortality among the caribou calves.

ed that predation by lynxes was what was causing the attrition of the caribou herds. The dead calves found in the field were the ones that had escaped after being bitten, only to die of septicemia. Biologists explored the wooded areas on the edges of the calving grounds and discovered the remains of other calves.

Like other feline predators, lynxes often grasp their prey by the neck and drag it under overhanging vegetation. The lynx apparently waits in the underbrush at the edge of the calving ground for an unwary calf to approach. The lynx dashes out, seizes the calf by the neck and drags it back into the brush. Male calves are more vulnerable to attack than females because the males tend to be more exploratory, often leaving their mother far behind to explore the forest edge. Female calves tend to stay closer to their mother. The adult female caribou in Newfoundland stands about 1.2 meters at the shoulder and weighs about 100 kilograms. This size is more than sufficient for her to drive off an attacking lynx weighing about 10 kilograms.

To test the hypothesis that the lynx was responsible for the decline of the caribou herds lynxes were removed by trapping from the area around two calving grounds. When this was done, there was a statistically significant increase in the fraction of calves that survived to join the herd. Caribou were then introduced onto several small islands along the coast of Newfoundland where there were no lynxes. As an experimental control caribou were also introduced into mainland habitats where the local caribou had become extinct but where lynxes were still present. On the small islands the caribou increased at a rate near the maximum calculated for the species. In the habitats where lynxes were present the rate of increase was half the maximum. In this same period the size of the main Newfoundland herd changed very little. It consisted of 6,100 animals in 1961 and 6,200 in 1966.

One of the places where the caribou had increased rapidly after their introduction was Jude Island. All the animals were introduced by the Newfoundland Wildlife Service under the direction of Eugene Mercer. Mercer was particularly closely involved in the work on Jude Island. Radio transmitters were attached to two lynxes and the animals were released on the island to observe their effect on the herd in an environment that had been free of predators. In the first season 65 percent of the calves were killed by lynxes, whereas in the preceding year all the calves had survived.

The work on Jude Island also revealed one reason the caribou dam stays near the body of her calf for several days. The movements of the lynxes, tracked by the radio transmissions, made it clear that after being driven away from its prey by the dam the lynx often remained hidden in the brush nearby waiting for her to leave. When the transmitter-bearing lynxes were removed from the island, the rate of calf survival immediately returned to its former high level.

After the experiment on Jude Island even the most skeptical observers were convinced that the lynx, which had never been reported to be a significant predator on caribou on the Canadian mainland, was the major factor in limiting the caribou population of Newfoundland. The timber wolf, once the chief constraint on the caribou population, was gone and the lynx had taken its place in the island's simple ecosystem.

If the lynx could regulate the caribou population, it seemed possible that it could also regulate the number of arctic hares. The two highland calving grounds where the rate of calf survival was highest were also the areas where arctic hares were still occasionally reported. On the other hand, it was possible that the snowshoe hares had reduced the population of native hares by competing with them for food. It was also possible that a change in the flora of the island had eliminated the indigenous hare's food supply.

To determine whether the increase of the arctic hares was limited by the amount of available food, two male and two female arctic hares were released on Brunette Island, about 16 kilometers south of Newfoundland in Fortune Bay. There were no mammalian predators or snowshoe hares on the island. Moreover, like much of Newfoundland, Brunette Island was a suitable habitat for arctic hares. Much of the island consists of subalpine and alpine zones, and the tundra hill summits are covered by mosses and ericaceous shrubs. Six reproductive seasons later there were 1,000 arctic hares on the island. Like the caribou, in the absence of lynxes the hares had multiplied at a rate close to the maximum for the species.

The release on Brunette Island proved that food was not a barrier to the growth of the arctic hare population. Arctic hares from Brunette were subsequently employed to test the other hypotheses; they were introduced to islands where there were snowshoe hares but no lynxes. As a control for the experiment arctic hares were introduced into areas of Newfoundland where there were both snowshoe hares and lynxes.

On the outlying islands where there were snowshoe hares but no lynxes the arctic hares survived and their population increased, although not as rapidly as it did on Brunette Island. In the Newfoundland habitats, where there were lynxes, none of the introductions succeeded. It was clear that the indigenous hare can withstand competition from the snowshoe hare but cannot withstand predation by the lynx, at least

in the lowland areas of Newfoundland.

The field tests confirmed that the lynx was limiting the abundance of both the caribou and the native hare on Newfoundland. From the release experiments and a study of the historical records of the island's fauna the outlines of an unusual dynamic system including the lynx, the caribou and the two species of hare began to emerge. For thousands of years before Newfoundland was settled by white men arctic hares, wolves and caribou coexisted on the windswept barrens and massifs of the island. In the early period the lynx was uncommon. Indeed, it was so rare that expert observers disagree on whether it should be considered native to the island at all.

The populations of wolves, caribou, lynxes and arctic hares undoubtedly fluctuated, but stabilizing mechanisms kept the small fluctuations from leading to the extinction of any one species. Then in 1864 snowshoe hares were brought to Newfoundland to help feed isolated fishermen on the west coast. The introduction of the new type of hare disrupted the delicate population dynamics that had prevailed among the indigenous species.

When an animal such as the snowshoe hare arrives in a new environment, it often tends to increase rapidly. After a few decades the introduced population reaches its highest level. By 1890 the entire island of Newfoundland was experiencing such a "high" in the snowshoe hare population. The density of the animals was so great that some inhabitants were able to trap 1,000 hares in a single season.

The initial high that follows the introduction of a species is generally followed by a decline as the inherent limits of the ecological niche occupied by the species begin to exert their effects. By about 1915 the snowshoe hare population had begun to decline rapidly.

In much of Canada the lynx is the chief predator on the snowshoe hare. As the snowshoe hare population was increasing rapidly in Newfoundland toward the end of the 19th century the lynx population was also increasing, although the curve of the lynx increase lagged behind that of the hares. This is a typical pattern in areas where the lynx and the snowshoe hare form an obligatory prey-predator pair. In such areas the curves representing the two populations take a regular sinusoidal form. The snowshoe hare population increases until it is limited by the food supply or by an intrinsic biological mechanism. The hare population then declines because its rate of reproduction has decreased. With fewer hares present the reproductive rate of the lynx also decreases and the lynx population decreases. In the absence of external perturbations the complete cycle takes from nine to 11 years.

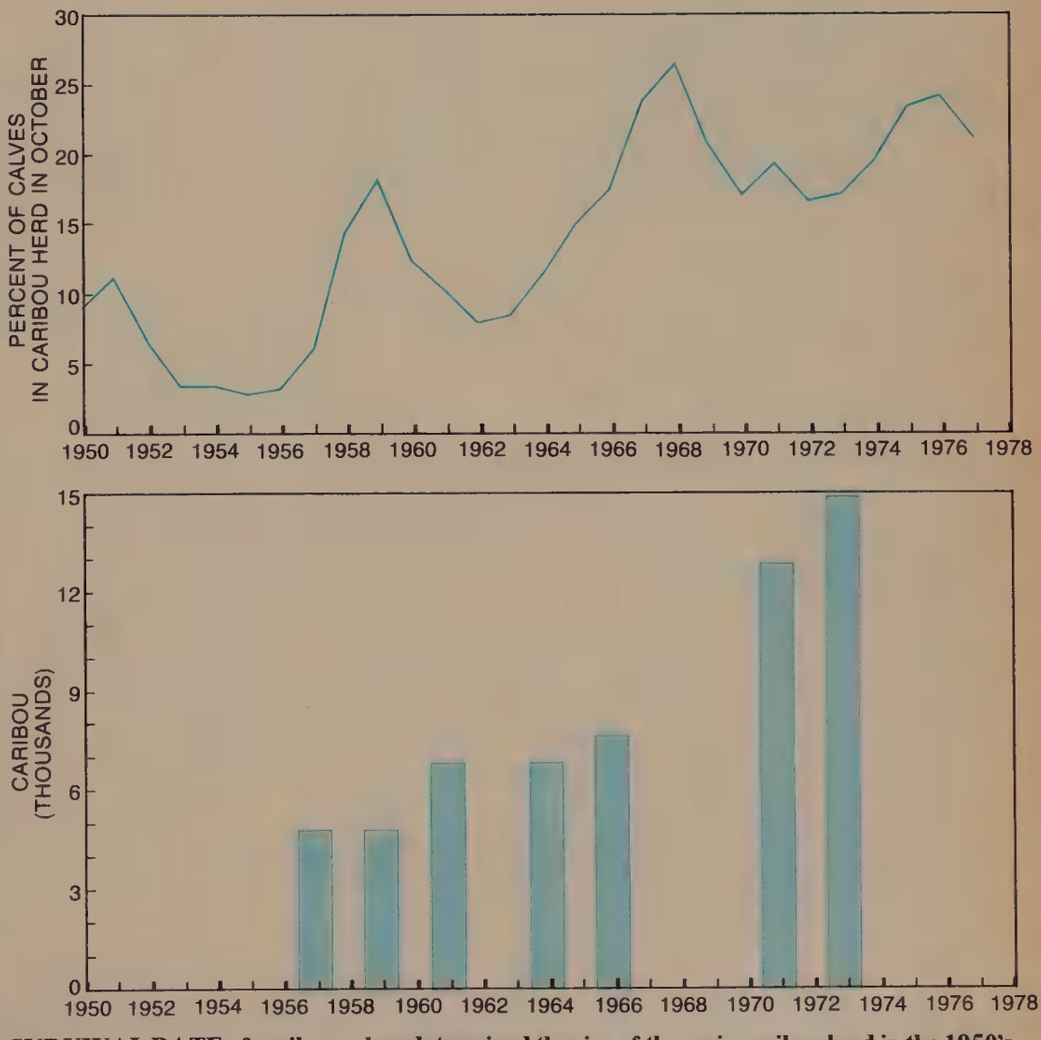
By the early years of this century the lynx had spread across Newfoundland from the west coast. Lynxes were reported within the city limits of St. John's, the capital on the Avalon Peninsula, which extends from the southeast coast. A bill to exterminate the lynx was proposed in the Newfoundland Assembly in 1904, which suggests how rapidly the lynx had increased. The lynx population was ultimately regulated, however, not by human factors but by ecological ones.

The snowshoe hare population remained at a peak from 1896 to 1915. (The oscillations of the hare and lynx populations had not yet become established.) When the hares crashed, the abundant lynxes were left without their main food item. Lynxes are resourceful scavengers and flexible predators. When food is particularly short, adult lynxes stop breeding but rarely starve. Instead they switch to alternative wild prey species, to domestic animals or even to ocean refuse deposited on the shore.

In Newfoundland the only alternative wild prey were mice, arctic hares and caribou calves. The lynxes presumably began to stalk caribou calves and the few available arctic hares. The new prey were not abundant enough to allow the lynxes to continue to increase or maintain their numbers, but they did mitigate the loss of the chief source of food.

At about the time lynxes began to take large numbers of caribou calves the death rate among the adult caribou rose for a quite different reason. The Newfoundland railroad, completed in about 1900, intersects the path along which the island's main caribou herd once migrated south in the fall. From 1911 to 1925 hunters, exploiting the railroad as a convenient means of pursuing the animals, began harvesting the herd in excess. In each of the years 1911, 1912, 1914 and 1915 hunters killed more than 5,000 caribou.

Because the high attrition among adult caribou came when the lynxes were beginning to prey on the calves, there were few new recruits to replenish the losses caused by the hunters. In a few years the 40,000 caribou of 1900 were all but wiped out. In the two dec-



SURVIVAL RATE of caribou calves determined the size of the main caribou herd in the 1950's and 1960's. The curve in the upper panel shows the fraction of the main herd consisting of calves in October. The fraction corresponds to the number of calves that survived the spring and summer. The peaks in calf survival in 1951, 1959, 1968 and 1976 coincided with snowshoe hare peaks on the island. When the snowshoe hare population crashed, calf survival declined. The bars in the lower panel show the total size of the herd in the interior of Newfoundland in the years for which data are available. When the prey-switching cycle was most intense, the size of the herd was determined by calf survival. Because of a decrease in the number of lynxes and the introduction of new prey species, the intensity of the cycle has recently diminished.

ades required for the catastrophic reduction of the caribou herd the snowshoe hare population had begun to fluctuate in the cyclical pattern observed in other parts of Canada. Peaks in the number of snowshoe hares were noted in 1920, 1931, 1940–1943, 1951–1952, 1959–1960, 1969 and 1976.

At the highs in the snowshoe hare cycle the lynxes switched back to their main prey and the rates of survival among caribou calves increased. Historical records show the caribou herds grew rapidly in about 1940, 1950 and 1960. During the 1960 snowshoe hare peak the chief caribou herds that were counted increased from 4,800 animals in 1959 to 6,100 in 1961. Thus the predatory adaptation of the lynx meant that the populations of both caribou and arctic hares were tightly coupled to the snowshoe hare population.

In recent decades the cycles of caribou-calf survival have become less pronounced; even at the low points in the cycle a larger fraction of calves survive than was the case when the cycle was at its most intense. This damping has occurred partly because the lynx population on the island has declined as a result of another cyclical phenomenon: women's fashions.

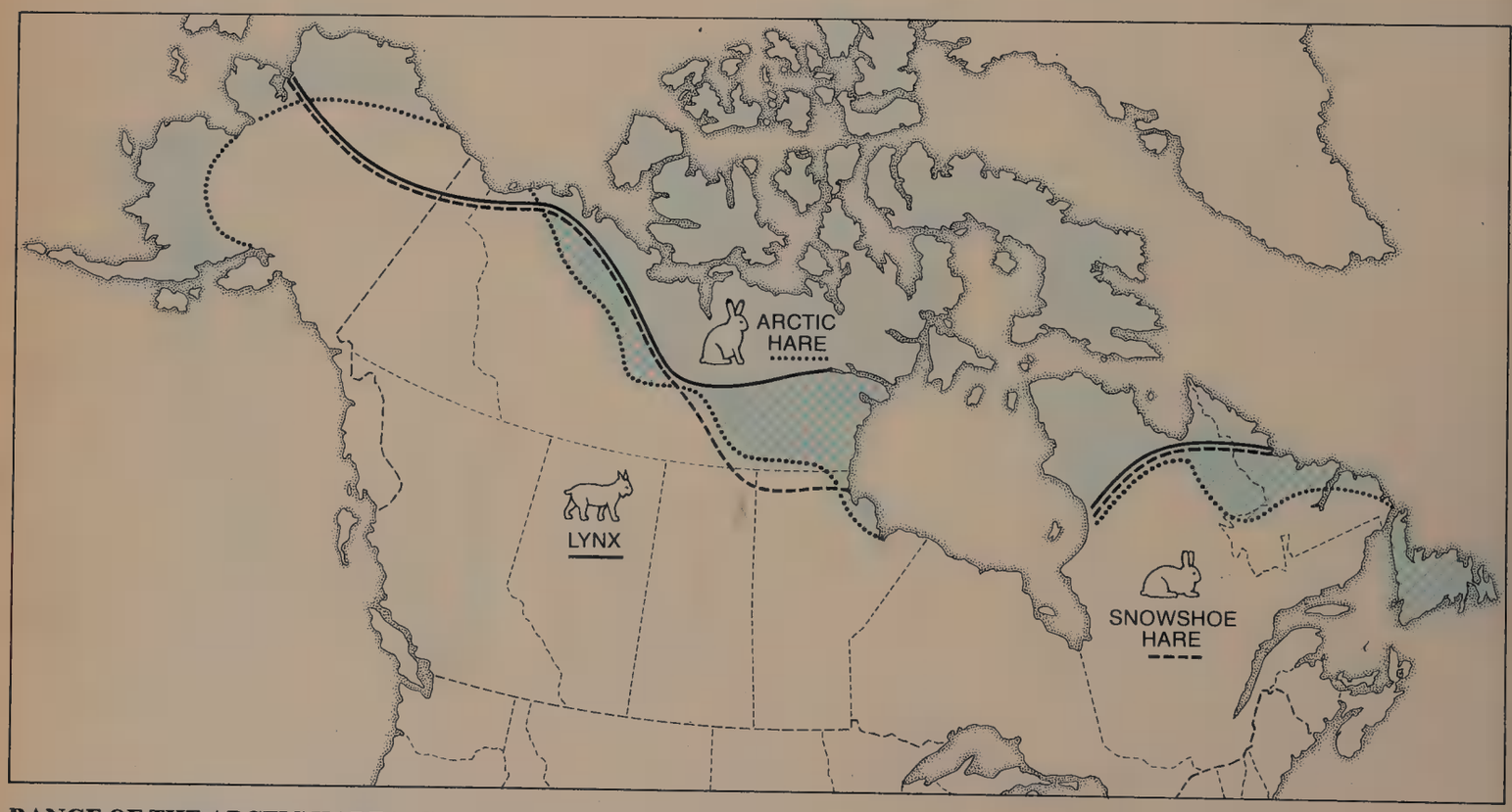
Long-haired animal pelts have recently become fashionable again after a hiatus. Trapping the lynx has therefore become economically rewarding, and the density of the lynx population in Newfoundland has dropped. Moreover, several new types of prey have been imported to Newfoundland, including the ruffed grouse and the spruce grouse. The island's food chain has become more complex, contributing an additional damping effect to the oscillations of predator populations.

Much of the effect the lynx has had on the caribou in Newfoundland stems from the fact that the caribou has adapted over a long period to predation by the wolf. The strategies that worked well against the wolf, however, were actually a handicap against a new predator with different hunting habits. For example, the large caribou aggregations that form after calving in the spring give the caribou a substantial advantage in protecting the calves from wolves. Wolves often hunt by day and stalk their prey in the open. When the caribou herd sights a wolf, its main defensive strategy is high-speed flight. The size of the herd at calving time enhances vigilance: it increases the probability that one female will see the wolf coming from afar. Moreover, the eye of the caribou is well adapted to detecting motion by day, a valuable asset against a diurnal predator moving in the open. By the same token, the unobstructed terrain of the calving ground makes it easier to detect a stalking wolf.

This combination of physiological and behavioral characteristics, however, is of little value when the predator is a lynx. The lynx hunts mostly at night and ambushes its prey rather than bringing it down in full-speed chase. Lynxes often crouch and freeze during their approach; this neutralizes the defenses of the caribou, whose eyes are better adapted to detecting motion than to detecting form. Caribou generally do not react with alarm to a man standing still, and they may be similarly unreactive to a crouched lynx.

When a predator is stalking a herd of herbivores, it often selects a particular animal to attack because that animal differs in some way from the others in the herd. Wolves choose young, sick, old or crippled caribou; almost any kind of conspicuous difference can make a caribou attractive to a wolf. With lynxes, however, this kind of selection is focused exclusively on the young. For a lynx the most conspicuous target is an active, curious calf at the edge of the calving ground near the fringe of vegetation that serves the predator as stalking cover.

It may be that predation by lynxes in the absence of wolves has pointed a new direction for the evolution of the Newfoundland caribou. Natural selection may currently favor the dispersal of calves rather than their concentration on the open calving ground, which makes them an obvious target. Moreover, the bond between the dam and the calf may be strengthened. A calf that



RANGE OF THE ARCTIC HARE in Canada (*color*) could be limited in part by predator-prey relations. The southern boundary of the range of the arctic hare coincides with the northern boundary of the range of the lynx: the range of the arctic hare overlaps only slightly that of the lynx (*hatched area*). The range of the lynx and the range of the snowshoe hare, however, largely overlap. The lynx and the snowshoe hare have a low foot loading: the pressure exerted on the ground by the feet when the animal runs. Thus they are well equipped for running on the soft snow of the boreal forest. The arctic hare has a higher foot loading for running on the hard snow of the tundra. In the boreal forest the arctic hare flounders and is vulnerable to the lynx. In the absence of lynxes the arctic hare might range farther south.

statements in application programs.

The non-navigational nature of the relational data base model endows QBE and SQL with extreme flexibility. Since there are no predefined information pathways to negotiate, the user is free to make all manner of ad hoc queries—an essential feature for applications where information needs change rapidly.

One of the most important IBM innovations in relational data base technology was a compiler approach to execute SQL statements. Replacing an interpretive approach, this compilation technique reduces the overhead cost of implementing the SQL language by using a precompiler to generate a tailored data access routine before execution time. The access routine, because it is tailored to one specific program, and is reusable, runs much more efficiently than a generalized interpreter.

In addition, the compilation technique uses a very sophisticated optimizer, which chooses economical access paths to the data. The compiler approach allows data base query in a high-level, easy-to-use language, yet also provides efficient program execution.

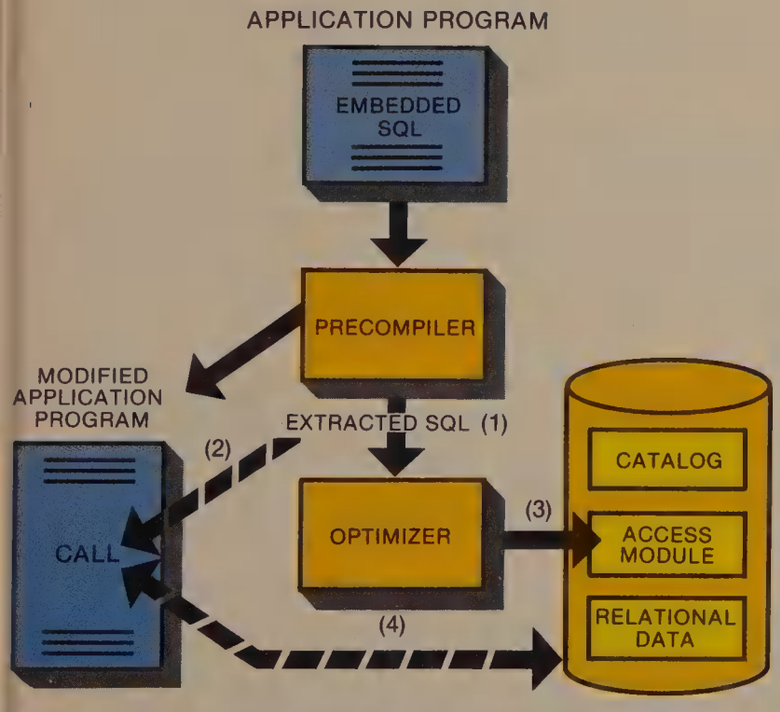


Figure 3. *The compiler approach is the key to IBM's efficient execution of SQL (very-high-level relational data base language) statements. This diagram illustrates the execution of application programs with embedded SQL statements. Programs are first processed by a precompiler, which extracts SQL statements from the application program (1). The precompiler also replaces the SQL statement in the host program with a CALL to the access routine (2). By very sophisticated analysis of available paths to the data, the optimizer chooses an economical path for the specific SQL statement, which is implemented as an access module (3). When the programs are executed, all the access modules for that program are loaded to provide targets for the modified CALLS (4).*

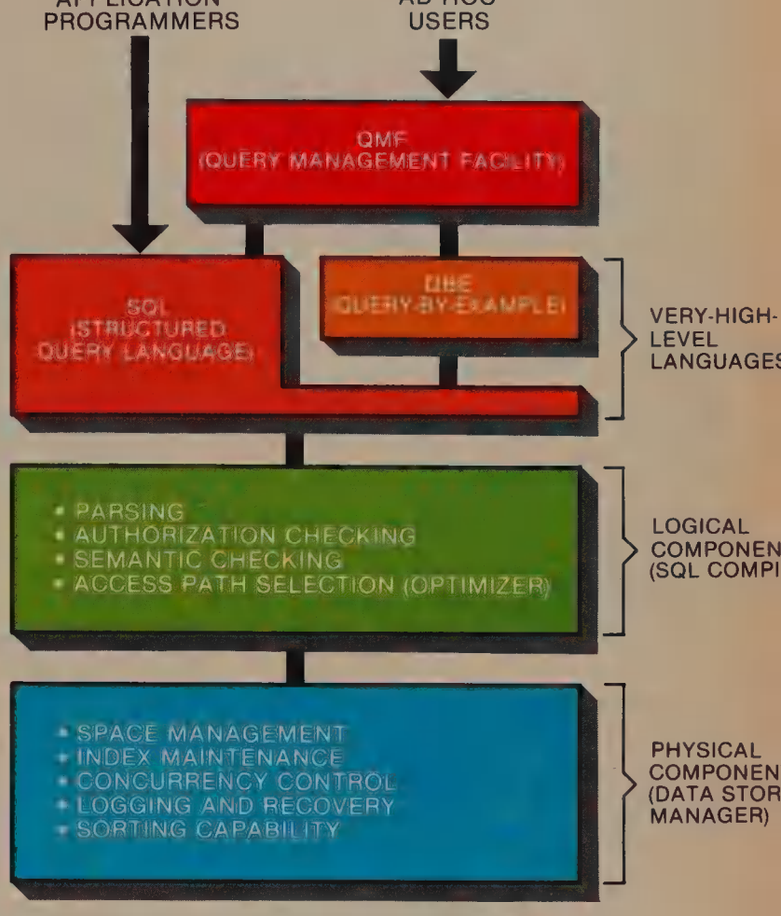


Figure 4. *This generalized architecture is the basis for IBM's relational data base products. It enables different types of users to access data easily, and yet is designed to handle complex programming tasks efficiently while providing the full function of a data base management system.*

Many scientists and programmers throughout IBM contributed to the development of relational data base technology, and researchers continue to explore future applications for the office environment and network users. These contributions are only part of IBM's continuing commitment to research, development, and engineering.

IBM®

For free additional information on relational data base, please write: IBM Corporation, Dept. 813E/3N35 Old Orchard Rd., Armonk, N.Y. 10504

stays close to a protective mother is likelier to survive predation by a lynx than a calf that can run fast, even though the calf that can run fast is better able to escape a wolf.

If the caribou adapts, its rate of survival could increase rapidly, because the reciprocal adaptation of the lynx must operate within narrow limits. The lynx cannot depend only on the caribou for food; it depends primarily on the snowshoe hare. Any change in its hunting behavior cannot alter the pattern that makes it an effective hunter of hares. A new evolutionary relation between the lynx and the caribou should therefore favor the caribou.

Like the predation of the lynx on the caribou, the relation between the lynx and the arctic hare has not been established over a long period of close coevolution. The two species have mostly separate domains. On a map of Canada the southern boundary of the arctic hare's range coincides closely with the northern boundary of the range of the lynx. The territories of the lynx and the snowshoe hare, on the other hand, are mostly overlapping: the predator ranges about as far north as the prey.

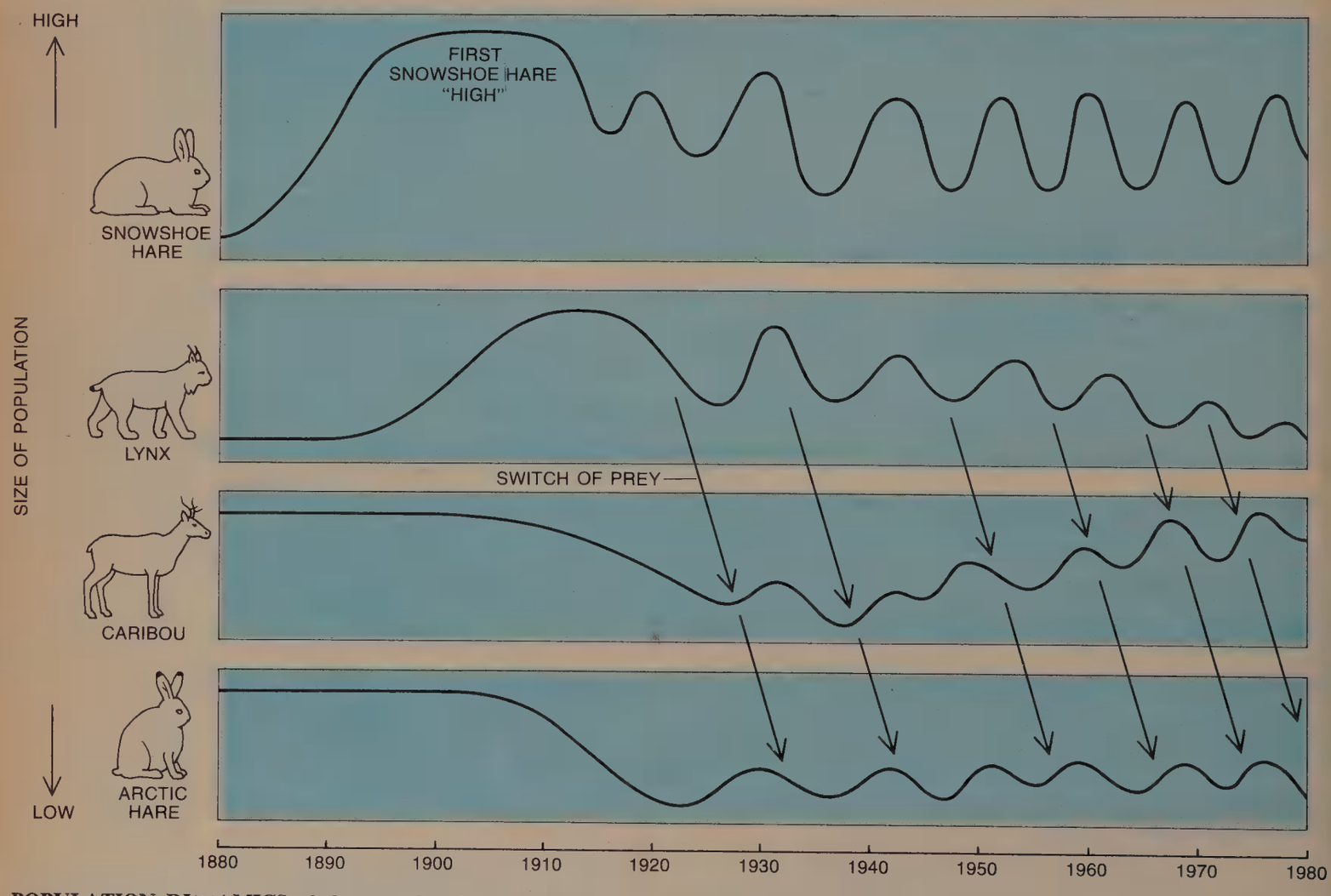
Most ecologists would explain the distribution of the three species according to the concept of biomes, or life zones: the arctic hare lives in the tundra biome, whereas the lynx and the snowshoe hare live in the boreal-forest biome. Predator-prey relations, however, could also influence the range of the arctic hare. It could be that in the absence of lynxes the arctic hare in Canada would range somewhat to the south of its current domain.

The success of the lynx in stalking the two species of hare depends on, among other things, the vegetation and the snow conditions. In forest habitats the lynx stalks and ambushes its prey, and the snowshoe hare sits motionless on the alert. The outcome of an encounter between them depends mainly on the distance between them when the prey sights the predator and runs.

Mercer has pointed out that as the lynx and the snowshoe hare sprint through the loose snow of the boreal forest they are evenly matched in one critical respect. Both animals have a low foot loading: the pressure exerted by the feet on the ground when the animal runs. As a result they can run quickly through the forest without sinking into the soft snow.

The arctic hare, on the other hand, is adapted to the hard-packed snow of the northern tundra. It has a foot loading more than twice that of the snowshoe hare. In soft snow the arctic hare flounders, greatly reducing its chances of escaping a lynx. This does not imply that snow conditions alone determine the range of the arctic hare. It is the presence of soft snow and predators with a low foot loading that constitutes the fatal combination.

The switching of predators from one prey species to another is currently a particularly promising area of investigation in ecology. The switches help to explain how stability can be maintained in small and fragile ecosystems. When the primary prey declines, a resourceful predator can shift to a secondary prey. Through this mechanism the predator holds its place in a food web that otherwise would break down. Switching also removes predation pressure from a de-



POPULATION DYNAMICS of the snowshoe hare, the lynx, the caribou and the arctic hare suggest that the driving force of the prey-switching system is the intrinsic 10-year cycle of the snowshoe hare. The first peak in the snowshoe hare population in Newfoundland was the highest; subsequent peaks have been somewhat lower. Before the snowshoe hares came to the island lynxes were so rare that it is disputed whether the animal is a native species. When the snowshoe hare population grew, the number of lynxes increased rapidly. Although there were some 40,000 caribou in 1900, as a result of predation by lynxes and human hunters there may have been 1,000 or fewer by 1925. The arctic hare population is in a "predator pit": it continues to exist because it is so scarce. The species cannot get out of the pit. If the number of arctic hares increases, the lynxes quickly reduce it when they switch after the next cyclical decline of snowshoe hares.

creasing prey population and enables it to survive.

Even so, switching can result in the decreasing prey population's surviving only at a very low density. This is true of the arctic hare in Newfoundland. The high density of lynxes, which is maintained by the snowshoe hare, ensures that the arctic hare density on the island will remain low.

In such a system the arctic hare is said to be in the "predator pit." Another instance of a prey species in such a situation is the caribou in Ontario and British Columbia. In those provinces the caribou forms one leg of a triangular system that includes the wolf and the moose. In both areas the caribou coevolved with the wolf in the absence of the moose. In the 19th century the moose spread into the areas from the south and east. The addition of the moose to the food base caused the wolf population to increase. Caribou, however, are easier for wolves to hunt than moose are, and the wolf takes caribou whenever it can.

In parts of Ontario and British Columbia the caribou are in the predator pit. They persist because they are rare. If the caribou increase, wolves switch back to them as a source of food and the caribou numbers fall. They cannot escape from the pit as long as the moose population continues to support a high density of wolves.

In some instances the prey species in such a system might survive not only because its numbers are low but also because the population of the species is spread so thin that hunting the animals is inefficient in terms of energy. For example, the present carrying capacity of Newfoundland for the arctic hare is defined by the minimum area per hare that is needed to make meetings between lynxes and their prey quite infrequent.

Indeed, such meetings must be rare enough for the survival of young arctic hares to compensate for the mortality among the adults. If the number, and hence the density, of arctic hares increases, this condition will be violated. Predation by lynxes will quickly restore the equilibrium. If this is so, spatial relations could turn out to be a highly significant feature of the dynamics of predator-prey systems.

The most significant conclusion to emerge from the study of the fauna of Newfoundland, however, is the considerable and unpredictable effect of introducing a new species into a simple ecosystem. What inhabitant of the island in 1864 could have imagined that the importation of a few snowshoe hares as a source of food for hungry fishermen would 100 years later be the cause of cycles in the survival of caribou calves and limit the population of arctic hares? In interfering with insular and fragile natural systems man must always tread softly.



The input: Thousands of engineering graduates and skilled technicians from Georgia Tech, Emory University, Southern Tech, Georgia State and 5 vocational colleges. A helpful R&D center at Georgia Tech. One of the world's busiest airports to transport goods. Low real estate costs and taxes. The output: Over 150 high-tech firms. And counting. To access: Contact Roy B. Cooper, MACFED, P.O. Box 1740, Atlanta, GA 30301. Call (404) 577-3838.

ADVANTAGE ATLANTA

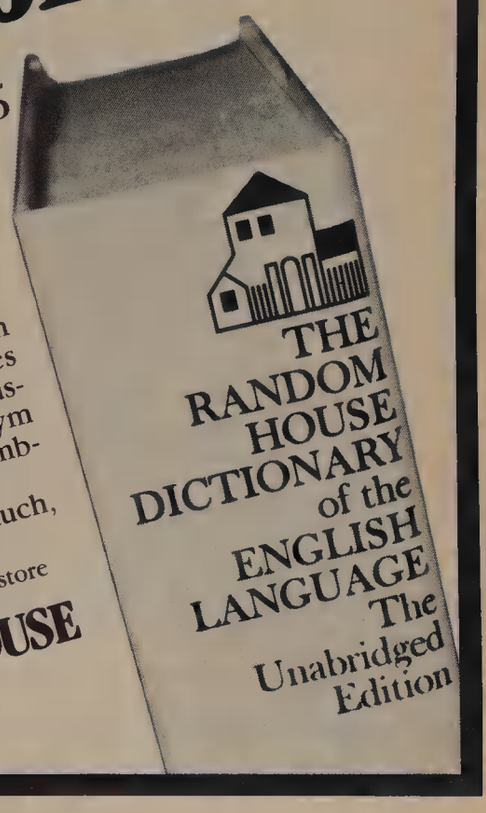
Clayton Cobb DeKalb Douglas Fulton Gwinnett

THE Answer Book The Unabridged!

- THE MOST AUTHORITATIVE DICTIONARY OF ITS KIND
- Comprehensive—more than 260,000 entries; 2,091 pages; large format 9" x 12" page size; 9 lbs 14 oz.
- Up-To-Date—with new words and terms
- Easy to Use—more than 50,000 example phrases and sentences; 2,000 illustrations; 10,000 synonym lists and studies; thumb-indexed
- Full-color ATLAS; much, much more

$49.95, now at your bookstore

RANDOM HOUSE



The Aerodynamics of Human-powered Land Vehicles

A bicycle and its rider are strongly impeded by their resistance to the flow of air. Aerodynamic stratagems have brought vehicles that can go 60 miles per hour on a level road without assistance

by Albert C. Gross, Chester R. Kyle and Douglas J. Malewicki

For decades the principles of aerodynamics have been applied with great success to improving the speed and efficiency of aircraft, automobiles, motorcycles and even competitive skiers and skaters. Vehicles powered by human energy, however, were virtually ignored until quite recently, which is strange in view of the fact that air resistance is by far the major retarding force affecting them. With a bicycle, for example, it accounts for more than 80 percent of the total force acting to slow the vehicle at speeds higher than 18 miles per hour. Here we undertake to explain this neglect and to show what attention to aerodynamics is beginning to do for the performance of human-powered land vehicles.

Looking first at the bicycle, one sees that it has remained almost the same in form for nearly a century. The Rover Safety Cycle, which was introduced in England in 1884, could easily pass for a modern bicycle; it lacks only a seat brace, which would have formed the modern diamond frame, and a few components such as brakes and multiple gears. Almost from the beginning the designers and users of bicycles recognized the importance of aerodynamics, but artificial constraints on design largely prevented the application of the necessary technology. It was as obvious then as it is now that wind forces at the bicycle-racing speed of from 20 to 30 m.p.h. are enormous.

Before 1900 the crouched posture of the bicycle racer had become common as a means of reducing air resistance. Another practice adopted before 1900 was to put a multiple-rider bicycle ahead of a single racer to shield him from the wind. In 1895 the Welsh wheelman Jimmy Michael rode 28.6 miles in one hour behind a four-man lead bicycle. In 1899 Charles ("Mile-a-Minute") Murphy of the U.S. gained international fame by pedaling one mile at 63.24 m.p.h. on a bicycle traveling behind a train of the Long Island Rail Road on a board path built for the occasion.

In 1912 Étienne Bunau-Varilla of France patented a streamlined enclosure for a bicycle and its rider that was inspired by the shape of the first dirigible balloons. Versions of this bicycle and its descendants set speed records in Europe from 1912 to 1933. In 1933 Marcel Berthet of France covered 31.06 miles in one hour riding a streamlined rig named the Vélodyne; his pace was more than 3 m.p.h. faster than anyone riding a standard bicycle had gone for one hour.

In the same year the French inventor Charles Mochet built a supine recumbent bicycle (with the rider pedaling while lying on his back) that he later streamlined. With a professional racer, François Faure, this "Vélocar" set a number of speed records between 1933 and 1938. Mochet and Faure hoped the records would be recognized by the Union Cycliste Internationale, the world governing body for bicycle racing. They were not.

Indeed, in 1938 the Union banned the use of aerodynamic devices and recumbent bicycles in racing; the rule is still in force. The ban has been a serious deterrent to the development of high-speed bicycles and is one of two major reasons the bicycle has remained nearly unchanged for so long. (The other reason is that in the developed countries the shift to the automobile has made the bicycle less important for transportation than it once was.)

By its ruling the Union essentially classified improvements in the aerodynamics of bicycles and other technological changes as "cheating." (It is perhaps fortunate that the Union was not active when a Scotch-Irish veterinary surgeon, John Boyd Dunlop, developed the pneumatic tire for bicycles in 1887, otherwise people might now be riding bicycles and possibly automobiles with solid steel wheels.) To its credit, however, the Union has gradually begun to relax its restrictions on changes in aerodynamics, although recumbents are still forbidden. Since 1976 skintight one-piece suits have become common in international bicycle racing. Streamlined helmets, teardrop cross sections for frame tubing, streamlined brake levers and other aerodynamically improved components have been allowed. In fact, technological change in all forms of human-powered vehicle is flourishing at a rate unmatched since the heyday of the bicycle in the 19th century.

This rapid change can be partly attributed to a series of events in California. In 1973 one of us (Kyle) and Jack H. Lambie, a consultant in aerodynamics who was working independently, built and tested the first two streamlined bicycles in the U.S. Unlike their predecessors, Kyle and Lambie actually measured the reduction in drag achieved by streamlining. They did so by conducting numerous coast-down tests, in which an unpowered vehicle is allowed to decelerate on a level surface. In this condition the deceleration of the vehicle is proportional to the total retarding forces acting on it; instruments measure either the speed or the deceleration. Kyle and Lambie, publishing their results independently, both concluded that the total drag forces on a bicycle could be reduced by more than 60 percent with a vertical, wing-shaped fairing that completely encloses the bicycle and the rider. (It was not until some two years later that either Kyle or Lambie learned that similar vehicles had been built earlier in Europe.)

In 1974 Ronald P. Skarin, an Olympic cyclist for the U.S., set five world speed records riding the Kyle streamlined bicycle at the Los Alamitos Naval Air Station. Because of this success, Kyle and

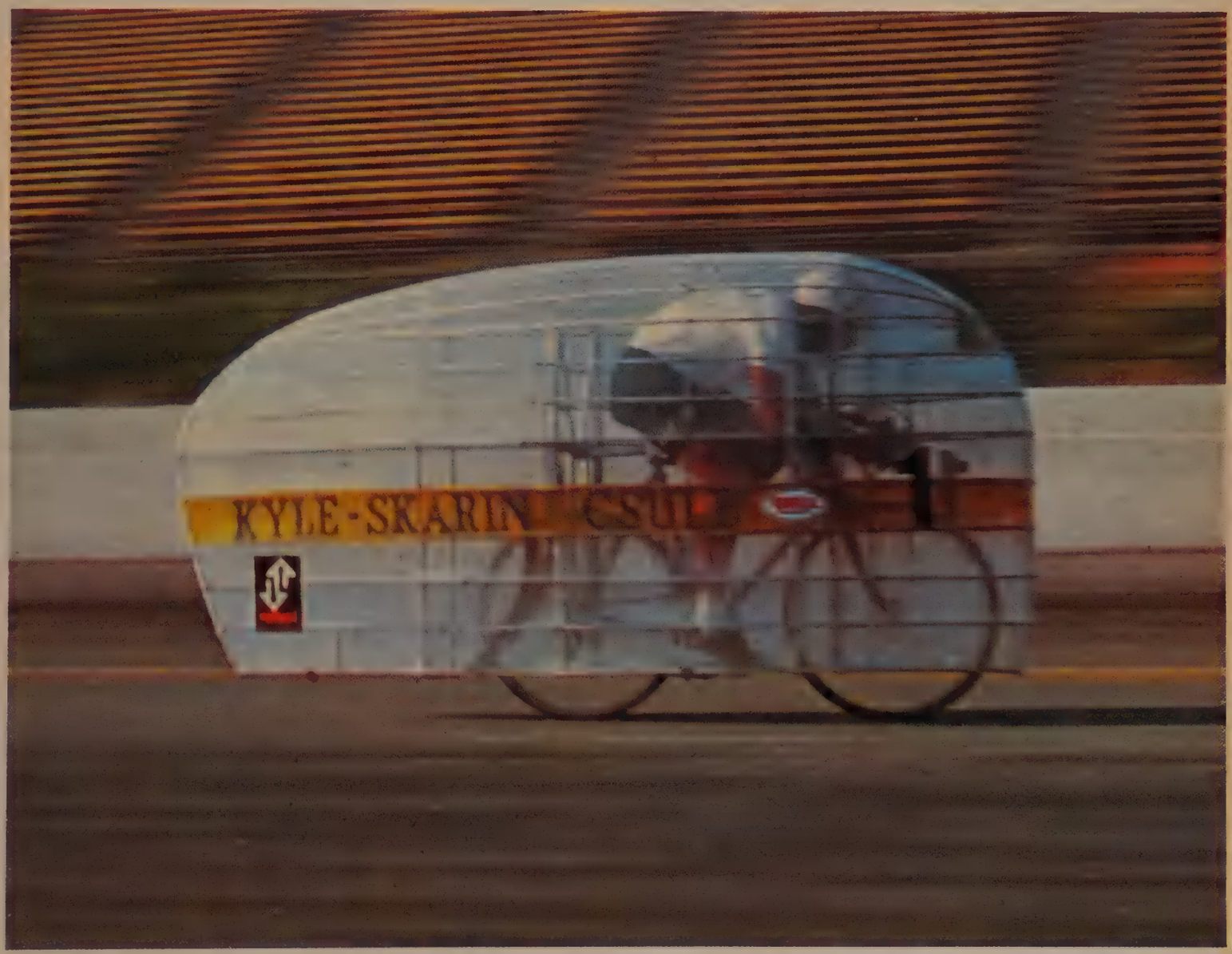
Lambie decided to organize a race for unrestricted human-powered vehicles. On April 5, 1975, at Irwindale, Calif., 14 distinctive vehicles competed in this historic first race. Many of them were recumbents, some with the rider pedaling supine (face up) and some with the rider prone (face down). Some were propelled by both hand and foot power. The winner at 44.87 m.p.h. was a streamlined tandem bicycle designed by Philip Norton, a high school teacher in Edgewood, Calif. The pedalers were Norton and Christopher Deaton, who is a skilled racing cyclist but not a world-class competitor. (The fastest an unaided standard racing bicycle has been ridden is 43.45 m.p.h., a record set in 1982 by Sergei Kopylov of the U.S.S.R., a cyclist of world class.)

Faced with the policy of the Union Cycliste Internationale against streamlining, the competitors in this race founded the International Human Powered Vehicle Association in 1976. Its purpose was to sanction competitions in which human-powered vehicles would be under no restrictions of design. Since then in dozens of races held in many countries the machines have become much more sophisticated and speeds have risen steadily. Four vehicles have broken the U.S. automobile speed limit of 55 m.p.h. (Each one received an honorary speeding ticket from the California Highway Patrol.) Among them is a third-generation streamlined quadricycle designed by Norton.

At present the world's fastest human-powered vehicle is the Vector Tandem, a gracefully streamlined two-person recumbent. It was built by a team headed by Allan A. Voigt, an engineer who as president of Versatron Research, Inc., primarily designs aerospace servomotors. (The pedalers ride supine and facing in opposite directions.) In 1980, with a flying start of about one mile of acceleration, it covered 200 meters along the track of the Ontario Motor Speedway in California at 62.92 m.p.h. Later that year the Vector Tandem averaged 50.5 m.p.h. for 40 miles on Interstate Route 5 between Stockton and Sacramento.

These extraordinary speeds are almost entirely the result of attention to aerodynamics. A cyclist traveling at 20 m.p.h. typically displaces approximately 1,000 pounds of air per minute. When the machine and the rider are not streamlined, they leave a substantial wake and exact a high cost in human energy.

Two types of aerodynamic drag affect the performance of a bicycle: pressure (or form) drag and skin-friction drag. Pressure drag results when the flow of air fails to follow the contours of the moving body. The separation changes the distribution of the air pressure on the body. If the separation takes place toward the rear of the body, the air pressure there becomes lower than it is on the forward surface, causing drag.



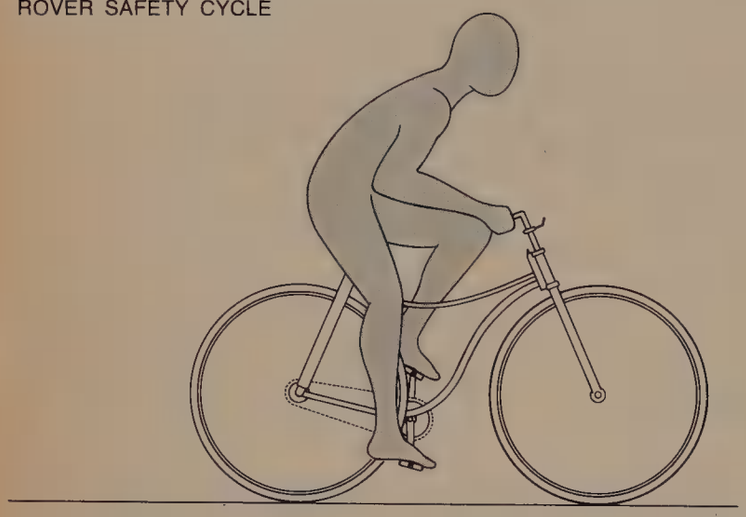
STREAMLINED RACING BICYCLE designed by one of the authors (Kyle) and ridden by Ronald P. Skarin, an Olympic cyclist for the U.S., is shown setting the world record of 31.88 miles per hour for one hour of pedaling from a standing start. The key to the performance was the streamlined fairing that reduced the aerodynamic resistance of the rider and the bicycle. Skarin established the new speed record in 1979 at the Ontario Motor Speedway in Ontario, Calif. Except for the fairing the vehicle was basically a standard racing bicycle.

Skin-friction drag results from the viscosity of the air. It is caused by the shearing forces generated in the boundary layer: the layer of air immediately next to the surface of the body.

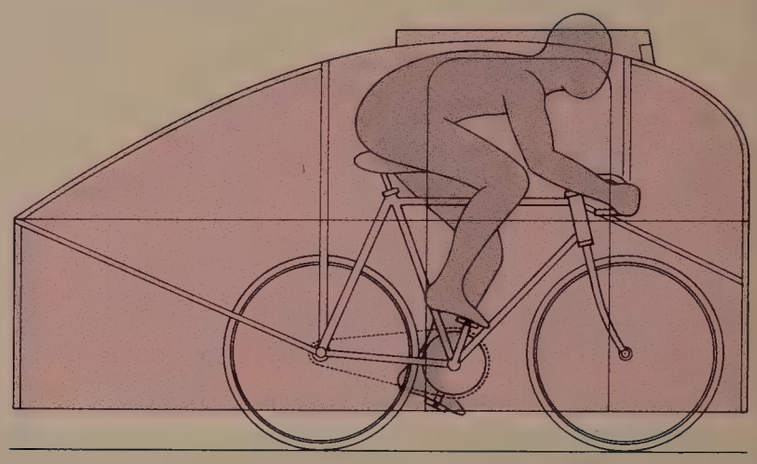
Blunt configurations such as the cylinders, spheres and other shapes found on a bicycle are aerodynamically inefficient because the airflow separates from the surfaces. Low-pressure regions form behind the objects, resulting in pressure drag hundreds of times greater than skin-friction drag. In contrast, air flows smoothly around a streamlined shape. The air closes in behind as the body passes. Pressure drag is greatly reduced and skin-friction drag becomes more important.

For the highest efficiency a vehicle should be designed to minimize the transfer of unrecoverable energy to the air by the two types of drag. At the present level of technology aerodynamic drag absorbs from 40 to 50 percent of the fuel energy consumed by an automobile or a truck at 55 m.p.h. Since the bicycle has lower power, weight and

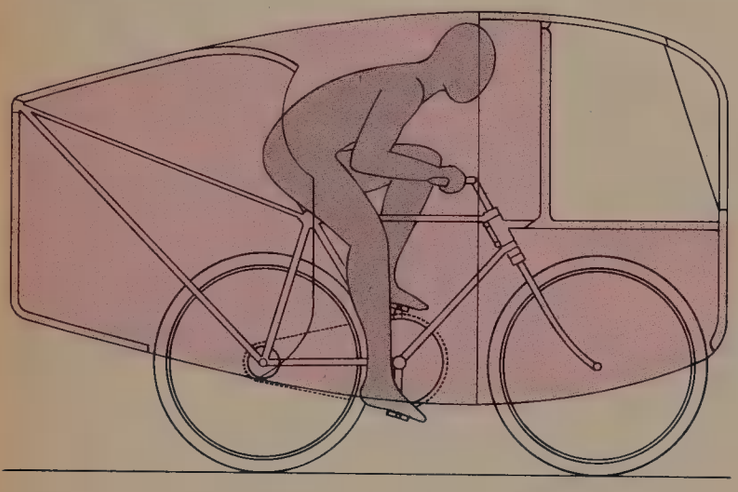
ROVER SAFETY CYCLE



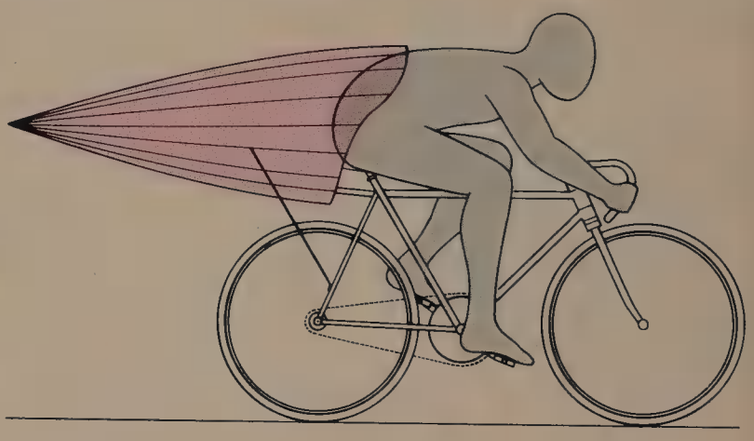
VÉLODYNE



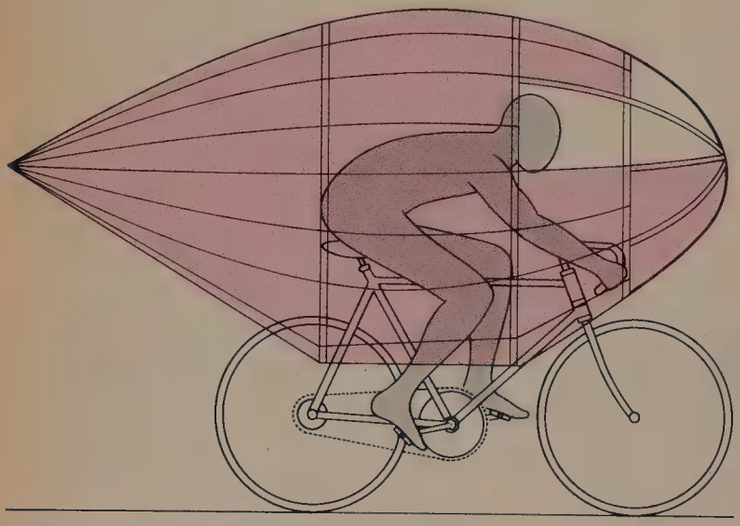
BUNAU-VARILLA DESIGN



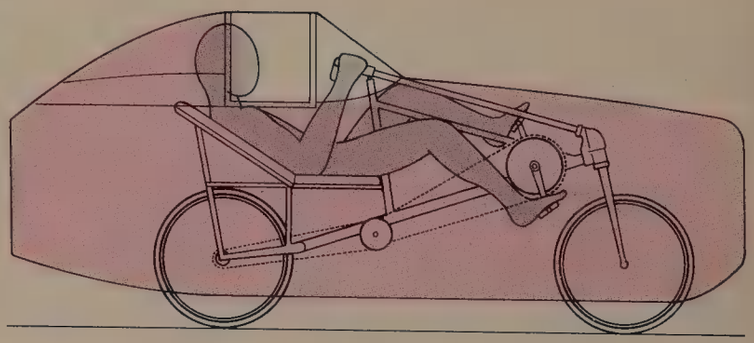
ROCKET



GORICKE



VÉLOCAR



EARLY IMPROVEMENTS in human-powered land vehicles resulted in the introduction of the Rover Safety Cycle in England in 1884. In 1912 and 1913 Étienne Bunau-Varilla of France obtained patents for a streamlined design; similar bicycles set many speed records. The Goricke was developed in Germany in 1914. The Vélodyne was ridden 31.06 miles in one hour (a new record) by Marcel Berthet of France in 1933. From the same year is the Rocket, designed by Oscar Egg. Another French vehicle, the Vélocar, set several speed records between 1933 and 1938. Most of the drawings are based on data from the Wolfgang Gronen Archive at Binningen in Switzerland.

rolling resistance and poor streamlining, aerodynamic drag accounts for an even higher percentage of the energy consumed at speeds above 10 m.p.h.

A term employed to describe the aerodynamic efficiency of a shape is the drag coefficient. An inefficient shape such as a sphere will have a drag coefficient of, say, 1.3, whereas a streamlined shape such as a teardrop will have one of less than .1. Hence an object of teardrop shape can move with less than a tenth of the loss of energy incurred by an object of cylindrical shape.

For land-transportation vehicles the aerodynamic resistance is almost directly proportional to the product of the frontal area and the drag coefficient. For convenience we call the product the effective frontal area. In discussing which of two vehicles has less aerodynamic drag it is not sufficient to compare drag coefficients; the size of the vehicle must also be taken into account. That is done in the concept of the effective frontal area. An ordinary bicycle and its rider will have an effective frontal area of from 3.4 to six square feet, whereas a streamlined human-powered vehicle can have one of less than .5 square foot.

The force of aerodynamic drag increases as the square of the velocity. Power is proportional to the product of drag force and velocity, so that the power necessary to drive an object through the air increases as the cube of the velocity. Hence a modest increase in speed requires an enormous increase in power. A cyclist who suddenly doubles his output of power when he is traveling at 20 m.p.h. will increase his speed to only about 26 m.p.h.

Conversely, reductions in aerodynamic drag affect speed less than one might think. If the air drag is cut in half at 20 m.p.h., a cyclist who does not change his power output will speed up to about 24.4 m.p.h. The reason is that the rolling resistance remains constant. If that resistance could be ignored, doubling the horsepower or reducing the effective frontal area by half would again get the speed up to about 26 m.p.h.

In sum, high speeds require extremely high aerodynamic efficiency. The Vector Tandem, receiving an input of slightly more than one horsepower from each of its two riders, attained a speed of 62.92 m.p.h. For a standard bicycle to achieve that speed would require more than 6 h.p. That level of power from a human rider is clearly impossible.

Designers and riders can reduce the aerodynamic drag on human-powered vehicles in three major ways. First, they can cut down the amount of energy wasted by the vehicle's interaction with the air. They do it by streamlining (reshaping the front and rear of blunt objects to minimize the pressure drag) and

WINNER OF AN AMERICAN BOOK AWARD

"THE BIOGRAPHY OF EINSTEIN HE HIMSELF WOULD HAVE LIKED THE BEST."*

The perfect book for anyone with an interest in science, written by Abraham Pais, an award-winning physicist who knew Einstein personally.

"A monument to sound scholarship and graceful style...Accurate, witty and clear as brook water, it is a work against which future scientific biographies will be measured."*
—The New York Times Book Review

"A sympathetic but clear-eyed view of his life and work.... a fine book."
—Scientific American

"An extraordinary biography of an extraordinary man."
—Christian Science Monitor

NOW IN PAPERBACK



'Subtle is the Lord'...

THE SCIENCE AND THE LIFE OF

Albert Einstein

by Abraham Pais

Illus. $25.00

OXFORD UNIVERSITY PRESS
200 Madison Avenue, New York NY 10016

Speak French like a diplomat!

What sort of people need to learn a foreign language as quickly and effectively as possible? *Foreign service personnel,* that's who.

Now you can learn to speak French just as these diplomatic personnel do — with the Foreign Service Institute's Basic French Course.

The U.S. Department of State has spent thousands of dollars developing this course. It's by far *the most effective way* to learn French at your own convenience and at your own pace.

The Basic French Course consists of a series of cassettes and an accompanying textbook. Simply follow the spoken and written instructions, listening and repeating. By the end of the course, you'll be learning and speaking entirely in French!

This course turns your cassette player into a "teaching machine." With its unique "pattern drill" learning method, you set your own pace — testing yourself, correcting errors, reinforcing accurate responses.

The FSI's Introductory Basic French Course comes in two parts, each shipped in a handsome library binder. Part A introduces the simpler forms of the language and a basic vocabulary. Part B presents more complex structures and additional vocabulary. Order either, or save 10% by ordering both:

☐ **Basic French, Part A.** 12 cassettes (15 hr.), and 194-p. text, $125.
☐ **Basic French, Part B.** 18 cassettes (25 hr.), and 290-p. text, $149.

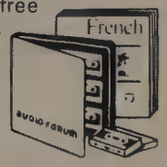
(Conn. and N.Y. residents add sales tax.)

TO ORDER BY PHONE, PLEASE CALL TOLL-FREE NUMBER: **1-800-243-1234.**

To order by mail, clip this ad and send with your name and address, and a check or money order — or charge to your credit card (AmEx, VISA, MasterCard, Diners) by enclosing card number, expiration date, and your signature.

The Foreign Service Institute's French course is unconditionally guaranteed. Try it for three weeks. If you're not convinced it's the fastest, easiest, most painless way to learn French, return it and we'll refund every penny you paid. Order today!

112 courses in 35 other languages also available. Write us for free catalog. Our 12th year.



Audio-Forum
Suite S28
On-the-Green,
Guilford, CT 06437
(203) 453-9794

AUDIO-FORUM®

All orders received by Dec. 12 will be delivered before the holidays!
Or visit our New York sales office 145 E. 49th St., New York, NY 10017 (212) 753-1783.

Embark on a wondrous journey across the universe of science

SCIENTIFIC AMERICAN LIBRARY

An incomparable collection of enduring value



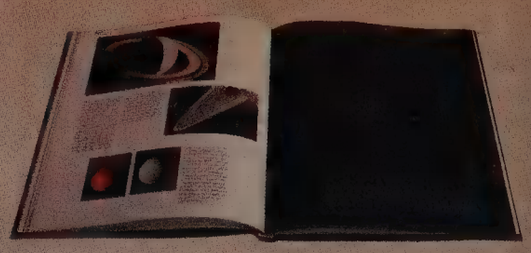
Out of *Scientific American's* collaboration with scientists in every field of endeavor, and after nearly a decade of planning, the SCIENTIFIC AMERICAN LIBRARY now offers a limitless resource for those intrigued by the wonder of science.

The SCIENTIFIC AMERICAN LIBRARY acquaints you with a world of understanding and shares with you the major ongoing advances at the frontiers of scientific investigation. Written by scientists who have made significant contributions, each volume rewards you not merely with information but with true understanding, as well as a command of the new concepts with which research proceeds.

As you receive your volumes you will see how they combine to make up a reference library of enduring value, one that embraces the unity of scientific understanding.

Share in this exciting enterprise with us. Simply fill in the attached card or coupon and we will send your splendid premier volume, POWERS OF TEN, to examine *free* for 15 days. You will be billed just $19.95 plus postage and handling for each volume, to be paid only *after* you have decided to add the volume to your personal library.

While you may cancel at any time with absolutely no notice, membership will bring you a cumulative library of scope and focus unparalleled in the literature of science.



YOUR PREMIER VOLUME

POWERS OF TEN

Philip and Phylis Morrison and the Office of Charles and Ray Eames

The essential survey of the universe from its largest dimension to its smallest

"If I were marooned on a desert island with just this one felicitous, densely informative book on all of science, I would count myself a lucky man. Having read it through, nothing on earth could prevent me from swimming off, any distance, to get another one like it."

LEWIS THOMAS, M. D.,
author of *The Lives of a Cell*

An atlas of the universe and a guide to its exploration, POWERS OF TEN takes you on an extraordinary adventure in magnitudes. This spectacular journey, with its sweeping changes in perspective, proceeds in ordered steps from the largest known dimension (10^{25} meters) to the smallest (10^{-16} meters).

You will find yourself abroad in a cosmos that in surprise and splender beggars the past creations of man's innocent imagination. In vivid color illustrations, POWERS OF TEN shows you the universe scaled from the human dimension. You will know better not only where you are but who you are.

The SCIENTIFIC AMERICAN LIBRARY offers POWERS OF TEN as its premier volume because it sets out the agenda for the books that follow.



MUSICAL SOUND

John R. Pierce

John Pierce shares with you his joy in the discovery that understanding from physics and mathematics enlarges and enhances one's appreciation of music. He traces the connection between Pythagoras and Galileo to the current musical innovators — Pierre Boulez, John Cage, and Edgard Varese — now working at the frontiers of sound and music.

John Pierce imbues readers with his confidence that rational inquiry into the highly subjective experience of music will open up new realms to the enjoyment of that experience.

Two 33⅓-rpm sound sheets demonstrate acoustic illusions.

SUBATOMIC PARTICLES

Steven Weinberg

As everyone knows, the atom, once considered irreducible, is composed of more fundamental particles: electrons, protons, and neutrons. But how do we know this?

In THE DISCOVERY OF SUBATOMIC PARTICLES we come to know the answer. With the author we reenact the experiments that disassembled the atom. We are there as J. J. Thomson, Ernest Rutherford, and James Chadwick discover, in turn, the electron, the proton, and the neutron. We gain a confident understanding of how the laws of physics govern each event we encounter.

HUMAN DIVERSITY

Richard C. Lewontin

The relevance of science to human values finds compelling demonstration in our understanding of the genetics of our species. In HUMAN DIVERSITY, Richard Lewontin shows that each of us differs from all others because of the interaction of genetic and environmental differences coupled with chance events that occur during our development.

Lewontin makes clear that the human genetic endowment confers a plasticity of psychic and physical development; each of us has the potential to develop any of a great diversity of identities within the human ambit.

FOSSILS

George Gaylord Simpson

In FOSSILS, noted paleontologist George Gaylord Simpson sets out the splendor of evidence upon which the theory and facts of evolution rest. He explains how "the primary record of the history of life is written in the successive strata of rocks as in the pages of a book. Fossils may be called the writing on those pages."

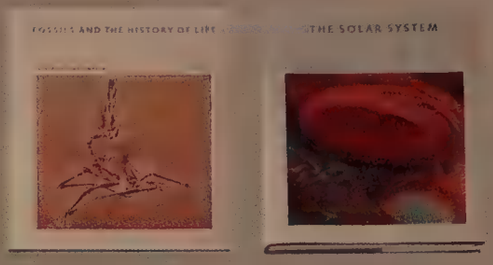
In his lifework and in this book, Simpson presents a picture of evolution very different from that of "nature red in tooth and claw" depicted in the first crude Darwinian popularizations and persistent still in the popular imagination.

THE SOLAR SYSTEM

Roman Smoluchowski

Standing on the moon, men have seen the Earth "small and blue and beautiful." Beyond doubt, the space expeditions of the U. S. and the U. S. S. R. must be reckoned as the most successful, as well as the most extravagant, ventures in the popularization of science.

In THE SOLAR SYSTEM, Roman Smoluchowski assembles a wealth of images from satellites and earthbound telescopes to give us a magnificent picture of the Earth and its cosmic neighborhood.



ON SIZE AND LIFE

Thomas A. McMahon and John Tyler Bonner

A biologist and an engineer bring their disciplines to bear in showing how nature obeys its own physical laws in fashioning living things. They demonstrate how size relates to shape and how this in turn controls the upper limit to the sizes of animals and plants.

Beyond the diversity of life, the authors show orderly ratios and relations that surprise and delight: the time for 50 percent of a mammal's growth is about three percent of a lifetime, and every mammal, no matter how long it can live, will take about 330 million breaths, sustaining 1.5 billion heartbeats.

PERCEPTION

Irvin Rock

The eye, we say, is a camera. Consider, however, that an object appears to us unchanging in size, shape, or orientation despite changes in our position with respect to it. How can we explain this constancy of perception when we know that the size, shape, and orientation of an image projected by the lens of the eye upon the retina must vary as we move?

In PERCEPTION, Irvin Rock explores the problems posed by both correct and illusory perception and examines the traditions of thought which inform contemporary investigations of the subject.

THE LIVING CELL

Christian de Duve

A trillion cells compose the human body, each cell about four orders of magnitude smaller than the world in reach of our unaided senses. In THE LIVING CELL, Christian de Duve brings the new understanding in the form of a wonderfully illustrated dynamic anatomy.

De Duve invites us to accompany him on an engrossing expedition through a world far removed from everyday experience yet vital to our very being. A bold but reassuring guide, de Duve puts the wisdom and knowledge gained from a lifetime of research at our service with his narrative, nontechnical, and often anecdotal style.

In addition to these volumes, leading scientists are at work on the exciting future books in the SCIENTIFIC AMERICAN LIBRARY. James Watson is writing on DNA, David Hubel on vision and the brain, Julian Schwinger on space and time. In the pages of the SCIENTIFIC AMERICAN LIBRARY, these Nobel Laureates and other distinguished scientists will explore the world as we know it.

Special Offer to Readers of *Scientific American*
Membership Reservation Form

To SCIENTIFIC AMERICAN LIBRARY
P. O. Box 646
Holmes, Pennsylvania 19043

Please enroll me as a member of the SCIENTIFIC AMERICAN LIBRARY. Send me the premier volume — POWERS OF TEN — by return mail. Thereafter, I am entitled to receive for inspection a new volume every two months, as it is published, for 15 day's *free examination.*

I need send no payment now. For each volume I keep I will be billed $19.95 plus postage and handling. This price is guaranteed for at least one year, regardless of any price increases asked of new members.

I understand that I may cancel my membership at any time.

NAME _____
(Please print clearly)

ADDRESS _____

CITY _____ STATE _____ ZIP _____

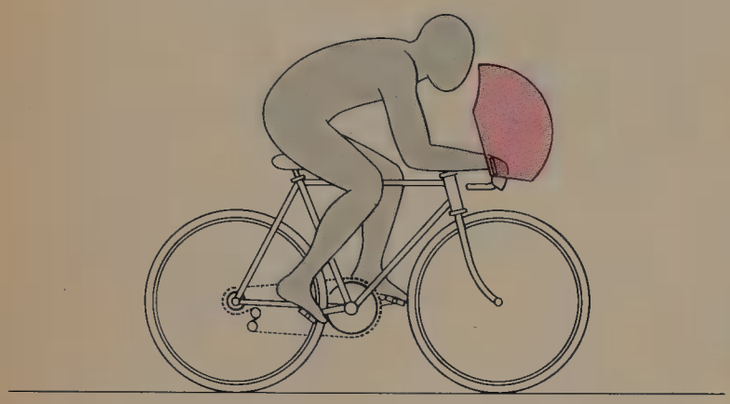
Canadian residents: U. S. dollars please

SSA123

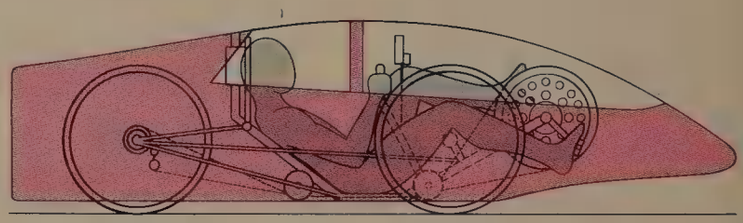
by smoothing rough surfaces to minimize skin-friction drag. Second, the amount of air encountered during each second of forward travel can be reduced. This is done by lowering the effective frontal area of the vehicle-rider combination. The same effect can be achieved by riding at higher altitudes. Third, the rider can find air moving in such a way that it provides a tail wind. Here the most effective approach is drafting, that is, riding closely in the wake of another vehicle.

At high altitudes the atmosphere is less dense and bicyclists encounter less air. In Mexico City (elevation 7,414 feet, where the air is only 80 percent as dense as it is at sea level) cycling records are from 3 to 5 percent faster than records made at lower altitudes. At La Paz, Bolivia (elevation 12,000 feet), sea-level

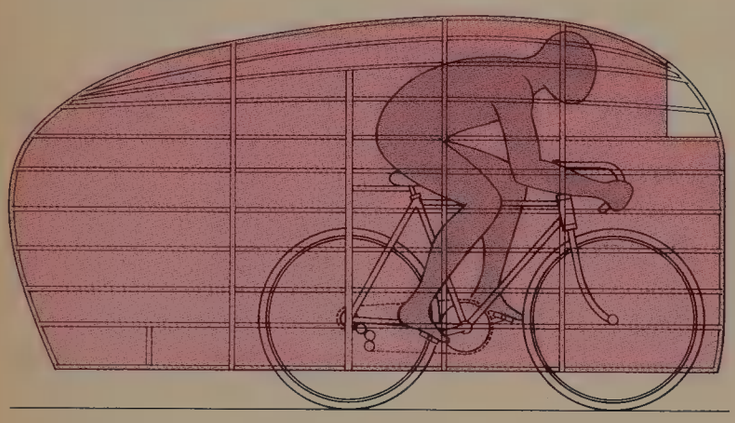
ZZIPPER



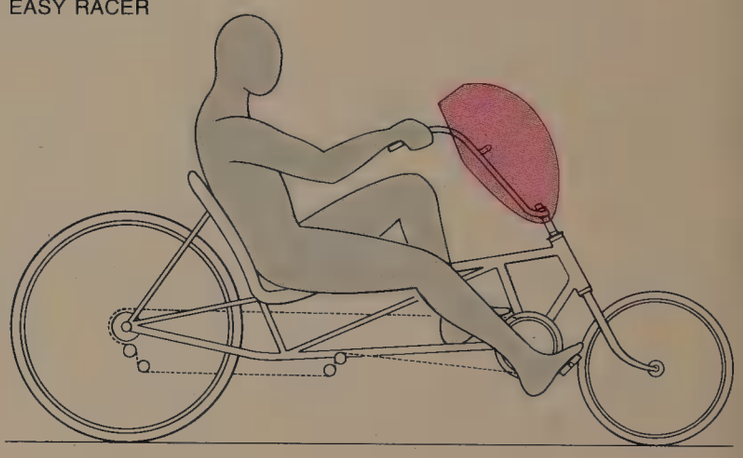
VECTOR SINGLE



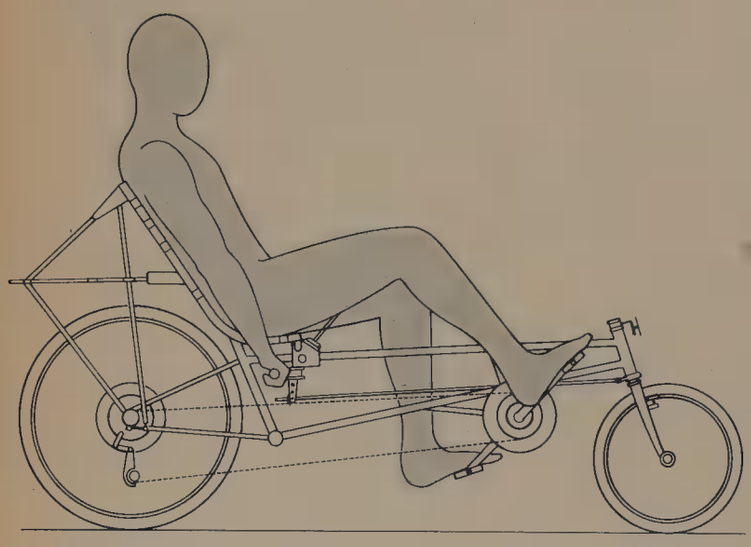
KYLE STREAMLINER



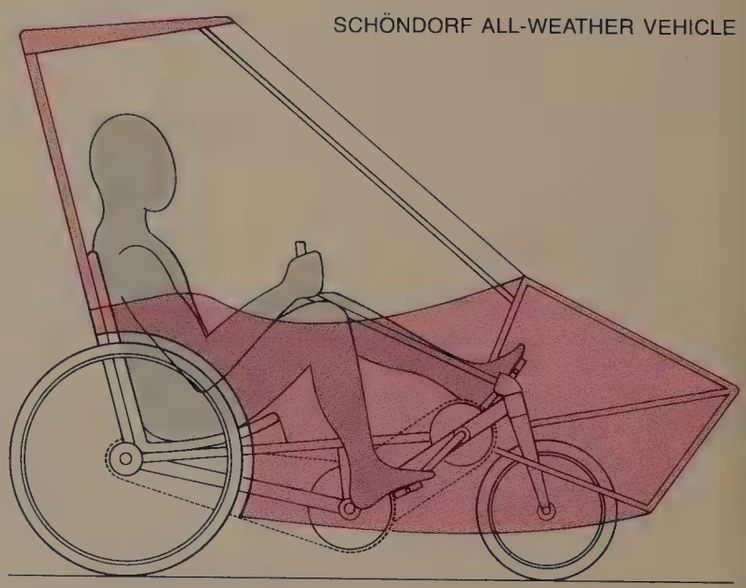
EASY RACER



AVATAR 2000



SCHÖNDORF ALL-WEATHER VEHICLE



MODERN HUMAN-POWERED VEHICLES make intensive use of streamlining to reduce the aerodynamic drag of the vehicle-rider combination. The simplest is the Zzipper, which is a partial fairing mounted in front of the rider. The Kyle Streamliner dates from 1973. A design that is meant for touring and commuting rather than for racing is the Avatar 2000; it utilizes the advantages of a recumbent position for the rider. The Vector Single, which has a full fairing, is theoretically capable of reaching almost 62 m.p.h. with an input of one horsepower from the rider. The Easy Racer is a recumbent designed mainly for touring or commuting, but it has also been raced. The last vehicle is one of the all-weather recumbents designed by Paul Schöndorf in Germany for elderly and handicapped people.

records could theoretically be improved by 14 percent. On the moon, where there is no atmosphere and only one-sixth the gravitational attraction, a suitably equipped bicyclist could theoretically ride at 238 m.p.h. with a very modest input of .1 h.p.

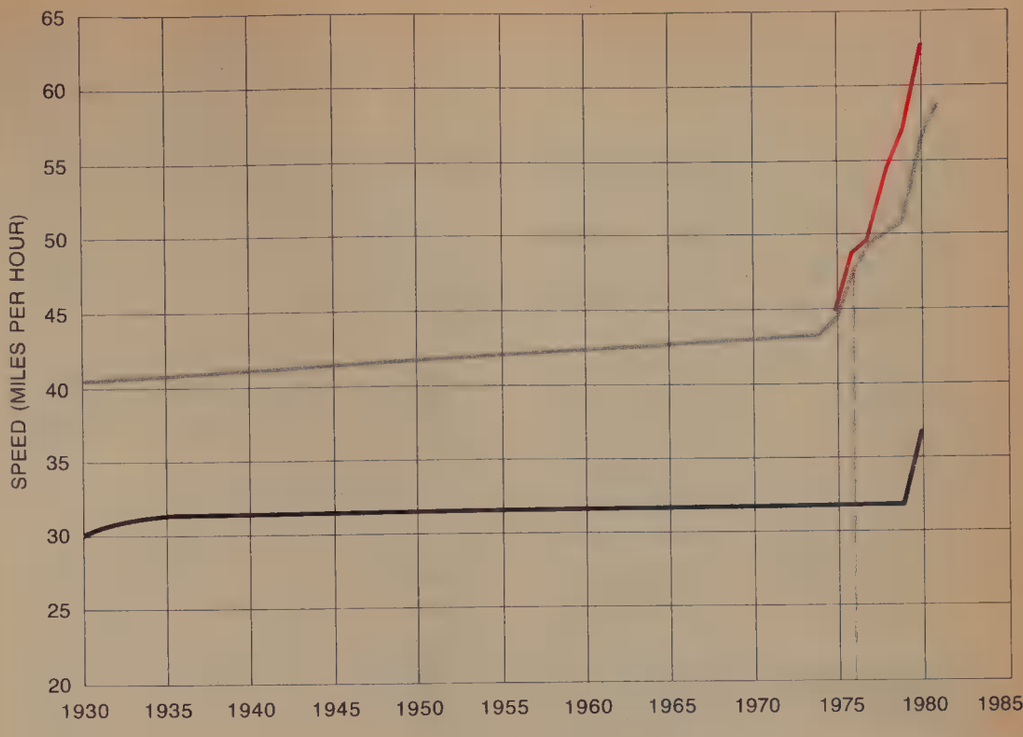
Analyzing the relation in which 80 percent of the power generated by a cyclist traveling on level ground at 18 m.p.h. goes to overcome air resistance, one finds that about 70 percent of the power consumption is due to the air's resistance to the rider and 30 percent to the air's resistance to the bicycle. This finding leads to the conclusion that to improve the performance of the standard bicycle one must first improve the aerodynamics of the rider.

For riders who race, the restrictions of the Union Cycliste Internationale leave little room for improvements beyond what has already been done in adopting the crouched position, the streamlined helmet, the skintight suit and the streamlining of components of the bicycle. As Voigt has calculated, even with a "perfect" bicycle (no aerodynamic drag on the machine at any speed and tires with no rolling resistance) the aerodynamic drag on the rider alone would severely hamper improvements in performance. According to Voigt, a crouched rider on a conventional racing bicycle could reach a maximum velocity of about 34 m.p.h. with a power input of 1 h.p. On a perfect bicycle the same rider making the same effort could achieve 38 m.p.h.

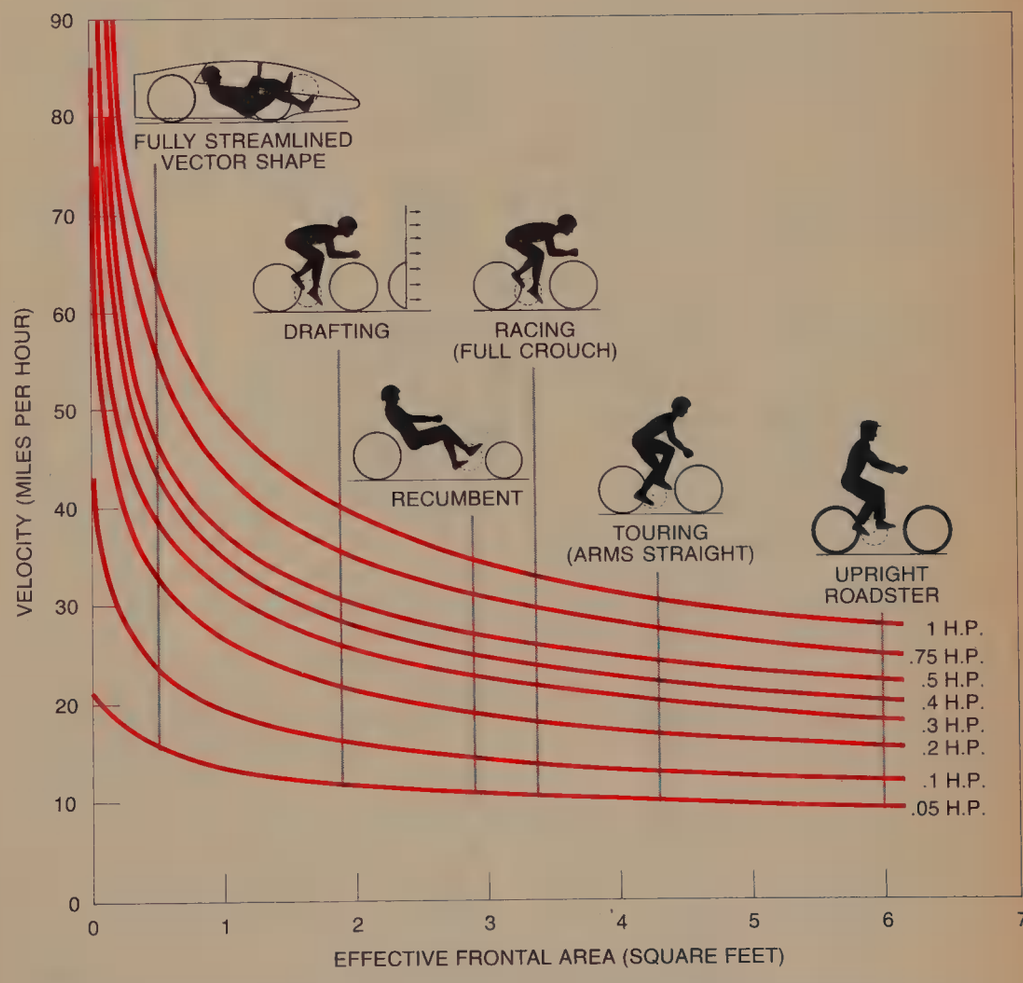
For the millions of noncompetitive cyclists who simply want a more efficient ride, several aerodynamic improvements are possible. They can be ranked in order of cost, beginning with the cheapest: a partial fairing such as the Zzipper, developed and manufactured by Glen Brown of Santa Cruz, Calif. It is a small, transparent, streamlined shield mounted in front of the rider. For about $60 a rider can lower the aerodynamic drag by about 20 percent, achieving a speed increase of some 2.5 m.p.h. for a 1-h.p. input.

Another effective way of reducing aerodynamic drag is to ride a recumbent bicycle. (The machine would cost several hundred dollars more than a basic touring bicycle.) The pioneers in this field are Gardner Martin of Freedom, Calif., designer of the Easy Racer, and David Gordon Wilson of the Massachusetts Institute of Technology, designer of the Avatar 2000. Because of the smaller frontal area presented by the recumbent rider, wind resistance decreases by 15 to 20 percent, resulting in about the same speed increase as is achieved by the Zzipper fairing.

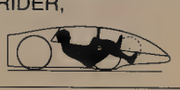
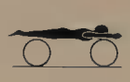
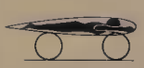
The recumbent bicycle offers other advantages. It is more comfortable to ride than a standard bicycle. In acci-



SPEED RECORDS of human-powered land vehicles have risen rapidly since the formation in 1976 of the International Human Powered Vehicle Association, which puts no restrictions on design. The time of the founding is indicated by the broken line. For many years before then the rules of the Union Cycliste Internationale, which banned streamlined vehicles from sanctioned bicycle competition, had kept speed records virtually unchanged. The curves represent records for multiple riders for 200 meters with a flying start (*color*), single riders under the same conditions (*gray*) and riders who pedaled for one hour at the maximum effort (*black*).



EFFECT OF STREAMLINING is to improve the performance of human-powered vehicles at all levels of power input. The upright roadster is the least streamlined vehicle, the Vector shape the most streamlined. Drafting means to follow closely behind another vehicle, here a bicycle. A good athlete can produce 1 h.p. for about 30 seconds, a healthy nonathlete for about 12 seconds. They can sustain an output of .4 and .1 h.p. respectively for about eight hours. The effective frontal area is the product of the drag coefficient and the projected frontal area.

	DESCRIPTION		FORCES AT 20 M.P.H. (POUNDS)	AERODYNAMIC DATA			ROLLING-RESISTANCE COEFFICIENT	LEVE
				DRAG COEFFICIENT	FRONTAL AREA (SQUARE FEET)	EFFECTIVE FRONTAL AREA (SQUARE FEET)		HORSE-POWER REQUIRED
STANDARD BICYCLES	BMX (YOUTH OFF-ROAD RACER)	30-LB. BIKE, 120-LB. RIDER, KNOBBY TIRES, 20-IN. DIA., 40 P.S.I. 	5.52 2.10	1.1	4.9	5.4	.014	146
	EUROPEAN UPRIGHT COMMUTER	40-LB. BIKE, 160-LB. RIDER, TIRES 27-IN. DIA., 40 P.S.I. 	6.14 1.20	1.1	5.5	6	.006	140
	TOURING (ARMS STRAIGHT)	25-LB. BIKE, 160-LB. RIDER, CLINCHER TIRES, 27-IN. DIA., 90 P.S.I. 	4.40 .83	1	4.3	4.3	.0045	100
	RACING (FULLY CROUCHED)	20-LB. BIKE, 160-LB. RIDER, SEWUP TIRES, 27-IN., DIA., 105 P.S.I. 	3.48 .54	.88	3.9	3.4	.003	77
IMPROVED MODELS	AERODYNAMIC COMPONENTS (FULLY CROUCHED)	20-LB. BIKE, 160-LB. RIDER, SEWUP TIRES, 27-IN. DIA., 105 P.S.I. 	3.27 .54	.83	3.9	3.2	.003	73
	PARTIAL FAIRING (ZZIPPER, CROUCHED)	21-LB. BIKE, 160-LB. RIDER, SEWUP TIRES, 27-IN. DIA., 105 P.S.I. 	2.97 .54	.70	4.1	2.9	.003	67
	RECUMBENT (EASY RACER)	27-LB. BIKE, 160-LB. RIDER, CLINCHER TIRES, 20-IN. FRONT, 27-IN. REAR, 90 P.S.I. 	2.97 .94	.77	3.8	2.9	.005	75
	TANDEM	42-LB. BIKE, TWO 160-LB. RIDERS, CLINCHER TIRES, 27-IN. DIA., 90 P.S.I. 	5.32 2.66 1.62 .81	1	5.2	5.2	.0045	66
	DRAFTING (CLOSELY FOLLOWING ANOTHER BICYCLE)	20-LB. BIKE, 160-LB. RIDER, SEWUP TIRES, 27-IN. DIA., 105 P.S.I. 	1.94 .54	.50	3.9	1.9	.003	47
RECORD HOLDERS	BLUE BELL (TWO WHEELS, ONE RIDER)	40-LB. BIKE, 160-LB. RIDER, SEWUP TIRES, 20-IN. FRONT, 27-IN. REAR, 105 P.S.I. 	.61 .8	.12	5	.6	.004	27
	KYLE (TWO WHEELS, TWO RIDERS)	52-LB. BIKE, TWO 160-LB. RIDERS, SEWUP TIRES 105 P.S.I. 	1.44 .72 1.12 .56	.2	7	1.4	.003	24
	VECTOR SINGLE (THREE WHEELS)	68-LB. BIKE, 160-LB. RIDER, SEWUP TIRES, 24-IN. FRONT, 27-IN. REAR 	.51 1.02	.11	4.56	.5	.0045	29
	VECTOR TANDEM (THREE WHEELS)	75-LB. BIKE, TWO 160 LB. RIDERS, SEWUP TIRES, 24 IN. DIA. 	.62 .31 1.78 .89	.13	4.7	.6	.0045	23
THEORETICAL LIMITS	PERFECT BIKE	NO ROLLING RESISTANCE, NO DRAG ON BIKE 	3.07 0	.8	3.8	3	0	59
	DRAGLESS RIDER	ROLLING RESISTANCE INCLUDES RIDER'S WEIGHT 	1.33 .81	1.1	1.2	1.3	.0045	41
	PERFECT RECUMBENT	DRAG ON RIDER ONLY 	.72 0	.6	1.2	.7	0	14
	PERFECT PRONE BIKE	DRAG ON SMALL BUT STRONG RIDER 	.51 0	.6	.8	.5	0	10
	PERFECT PRONE STREAMLINER		.07 0	.05	1.4	.07	0	1
	MOTOR PACING	42-LB., BIKE, 160-LB. RIDER, MOTORCYCLE ROAD-RACING TIRES, 70 P.S.I. 	0 1.21			VARIES WITH SPEED	.006	23
	MOON BIKE	25-LB. BIKE, 160-LB. RIDER, 15-LB. SPACE SUIT 	0 .15			0	.0045	3

GROUND, NO WIND		EFFECT OF HILLS	
ALL-DAY TOURING	MAXIMUM SPEED	STEADY SPEED UP	STEADY-SPEED COASTING
10.1	27.8	12.2	19.8
11.3	27.6	10.9	24
13.1	31.1	12.2	27.7
14.7	33.9	13	31.2
15	34.6	13	32.2
15.4	35.7	13.1	33.9
14.4	35.2	12.5	33.7
15.2	36.6	13	35.2
17.5	41	13.6	41.7
22.5	58.6	12.9	77.4
23.3	56.6	14	69.9
21.8	61.2	11.3	90.1
25.6	72.5	13	108.4
16.7	35.9	13.4	34.7
18.4	45.8	13.3	50.3
27.1	58.3	16.8	66.9
30.4	65.3	23.2	65.3
58.3	125.9	25.6	174.5
29.4	294	12.6	?
237.5	2,375	78.4	?

dents that do not involve an encounter with an automobile it is much safer, since the rider is closer to the ground (making falls less serious) and the feet are forward (making a head injury less likely in a fall). A problem is that a recumbent is hard to see on a road and so is perhaps more vulnerable to automobiles; the problem can be relieved somewhat by mounting on the vehicle a long, thin pole with a flag.

At the top of the expense ladder is a bicycle with a full fairing. The Vector Single, a one-rider version of the Vector Tandem, is the best example of a fully faired, enclosed, pedal-powered vehicle. (It is the machine portrayed on the cover of this issue of *Scientific American*.) According to Voigt, the vehicle is theoretically capable of reaching 61.7 m.p.h. with a 1-h.p. input, an increase of 28.2 m.p.h. over what has been done with a standard racing bicycle. A Vector Single costs about as much as a first-class racing bicycle.

In going up or down a hill a fully streamlined vehicle retains its advantage over a conventional bicycle. Although the Vector Single weighs about 80 pounds, compared with about 25 pounds for a standard bicycle, it can climb moderate hills as fast as or faster than the bicycle. With an input of .4 h.p. a bicycle can climb a 2.5 percent grade at about 16 m.p.h. and a 6 percent grade at about 11 m.p.h. With the same input the Vector can climb the two grades at 20.5 and 11 m.p.h. respectively.

Downhill the difference between the two machines is remarkable. The bicycle can descend a 2.5 percent grade at 29.5 m.p.h., the Vector at 54. On a 6 percent grade the bicycle can reach a speed of 39 m.p.h. and the Vector can exceed 100. Such potential speeds mean that if streamlined human-powered vehicles become common, careful attention must be given to the design of brakes and suspension and to the stability of the vehicle.

Since the aerodynamic drag force is proportional to the square of the relative velocity, head winds, tail winds and even crosswinds can drastically change both aerodynamic drag and the power requirements. For example, a bicyclist going at 18 m.p.h. in still air must increase his power output by 100 percent to maintain that speed against a head wind of 10 m.p.h. Usually a bicyclist confronting a head wind slows down and tries to maintain his customary leg force and pedaling cadence by shifting gears. This is one reason bicycles with multiple gears are desirable even for level country.

A tail wind makes the bicyclist go faster with his customary input of power. In general moving air will speed up or slow down a bicycle by about half the wind speed. When one bicyclist rides in the wake of another, the power requirements of the drafting rider are reduced by about 30 percent. The forward bicyclist creates an artificial tail wind.

The closer the rear bicycle follows the leader, the more pronounced the drafting effect is. One can think of the rear rider on a tandem bicycle as drafting extremely closely. Tandem riders use 20 percent less power per rider than two separate cyclists.

When the riders in a line of drafting bicyclists take turns in the lead position, the entire group can travel much faster than a single rider. In a pursuit race of 4,000 meters (almost 2.5 miles) a four-rider team can go about 4 m.p.h. faster than a single bicyclist. Typically a group of bicycle tourists of equal ability can travel from 1 m.p.h. to 3 m.p.h. faster than any rider alone. The larger the group is, the faster it should be able to travel (up to, say, a dozen riders).

Artificial winds created by passing automotive traffic can increase a bicyclist's speed from 1 m.p.h. to 3 m.p.h. for periods of about seven seconds. The larger the passing vehicle, the more substantial the effect. A steady stream of traffic can enable a bicyclist to sustain a speed from 3 m.p.h. to 6 m.p.h. higher than would otherwise be possible for a given energy input.

When a bicyclist rides directly in the wake of a motor vehicle, quite remarkable speeds can be attained. The practice is called motor pacing. On August 25, 1973, Allan V. Abbott, a physician in California, achieved a record of 138.674 m.p.h. motor pacing along a measured mile at the Bonneville Salt Flats in Utah. John Howard, a U.S. Olympic cyclist, is attempting to break Abbott's record and to achieve a motor-pacing speed in excess of 150 m.p.h.

Although the findings we have described are significant in their own right, one wonders if they will have any practical application beyond their effect on speed records. For a large fraction of the world's bicycle riders it seems unlikely that the work will have much immediate utility. For example, in the

PERFORMANCE OF HUMAN-POWERED VEHICLES is summarized. The numbers listed under forces for each vehicle represent air resistance and rolling resistance respectively. The five columns at the far right represent respectively the horsepower required at 20 m.p.h. as a percentage of the touring rider's performance; the all-day touring speed in miles per hour at an output of .1 h.p.; the maximum speed at an output of 1 h.p.; the steady speed in miles per hour up a 5 percent grade at an output of .4 h.p., and the coasting speed down the same grade.

many developing countries where the bicycle is the chief means of transportation most riders travel at about 7 m.p.h., often with a substantial load; aerodynamic drag becomes more important than other impediments to bicycle motion only at speeds above 10 m.p.h. Even here the work on aerodynamics makes a contribution. Without it designers would not know why they should largely ignore aerodynamics for slow-moving human-powered vehicles.

For bicycles intended for slow but sure progress it makes sense to decrease rolling resistance by improving tires and by paving roads. Designers should also reduce the bicycle's weight to facilitate climbing hills. The recent introduction of "mountain bikes" in the U.S. is a step in the direction of making lightweight bicycles durable enough for rugged or unpaved roads.

In several ways the knowledge gained by the recent research on the aerodynamics of human-powered vehicles can be directly useful. Although the standard bicycle is likely to be the predominant representative of the class for many more years because of its public acceptance, low cost, simplicity and mechanical reliability, it offers plenty of scope for innovation. For example, a light, simple and inexpensive front fairing will substantially improve the performance of the standard bicycle. The recumbent bicycle may come into greater use by commuters and tourists because of its efficiency and comfort.

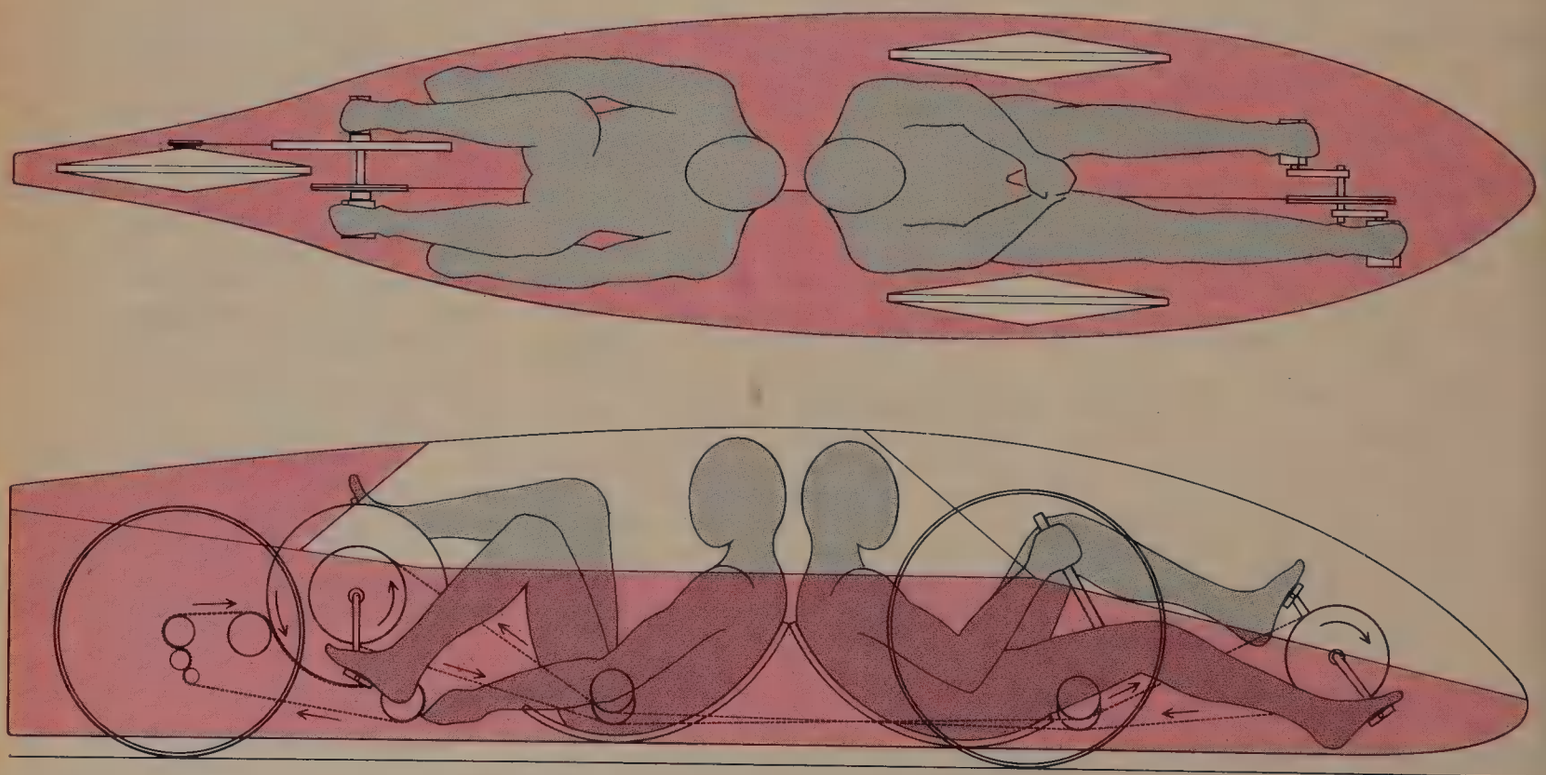
A further application of the technology would be to fit a recumbent with a small, lightweight motor of low horsepower. The motor would serve mainly as an aid in accelerating from a stop and in climbing hills. Fitted also with as much streamlining as would be consistent with the need for ventilation and stability, the machine would be a true moped. (The machines now sold under that name are not really motor-pedal vehicles but merely underpowered motorcycles.)

The recent research has inspired inventors to develop several special-purpose human-powered vehicles. Paul Schöndorf, a professor of engineering at the Fachhochschule in Cologne, has built a series of easily pedaled, all-weather recumbent tricycles for the elderly and the handicapped. Similar vehicles would serve well in retirement communities. Douglas Schwandt of the Veterans Administration's Rehabilitation Engineering Research and Development Center at Palo Alto, Calif., has built hand-cranked tricycles and bicycles for paraplegics. William Warner, a paraplegic who once held the record for hand-powered vehicles in the races sponsored by the International Human Powered Vehicle Association, says a disabled person can propel such a vehicle much faster than a standard wheelchair and thereby can gain a new sense of freedom and mobility. (The present record of 25.09 m.p.h. was set in 1981 by Ascher Williams of the Palo Alto rehabilitation center.)

In principle a fully enclosed, streamlined human-powered vehicle could be quite useful in transportation. A rider could travel at speeds of from 20 to 30 m.p.h. in all kinds of weather. As such vehicles are now designed, however, they would not serve on the open road. They lack adequate ventilation, visibility, maneuverability and such safety features as lights and windshield wipers. Most of them are not easy to get into or out of.

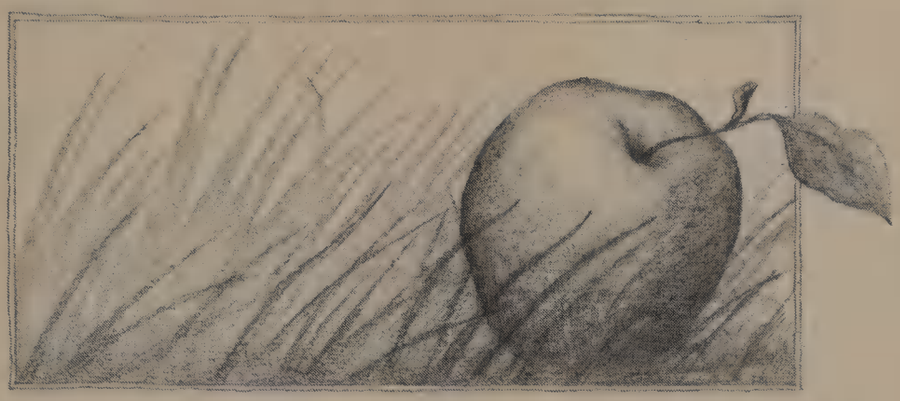
To produce a practical vehicle of this kind would require an investment and an engineering effort comparable to that made in producing a new automobile. Even then the pedaled vehicle would not be safe in traffic that included a large number of motor vehicles. One must conclude that a fully enclosed human-powered vehicle will not be a practical form of transportation until fuel shortages remove most motorized vehicles from the roads or until special roadways are built for pedaled machines.

Far likelier is the development of lighter and more fuel-efficient automobiles employing much of the technology we have described. One of us (Malewicki) has already built such a vehicle, a single-passenger machine weighing 230 pounds. It holds records for fuel economy at the freeway speed of 55 m.p.h. with a gasoline engine (157.2 miles per gallon) and with a diesel engine (156.3 m.p.g.). The diesel record was set on a trip from Los Angeles to Las Vegas, during which the average speed was 56.3 m.p.h. A trend toward such vehicles could help to extend fuel resources and ironically might postpone the time when the human-powered vehicle will have come fully into its own.



VECTOR TANDEM is shown in plan and elevation. It is a companion vehicle to the Vector Single portrayed on the cover of this issue of SCIENTIFIC AMERICAN. The Tandem, receiving an input of a bit more than 1 h.p. from each of its two riders, who are positioned back to back, set the speed record of 62.92 m.p.h. for 200 meters in 1980. (The riders had a flying start of more than one mile.) Traveling on an interstate highway in California later that year the Vector Tandem managed an average speed of 50.5 m.p.h. on a trip of 40 miles.

After the fall



Newton's apple has had a profound impact on modern physics. The goal of modern physics—describing and explaining nature's diverse and complex phenomena in a few basic laws—is realized in these landmark books from W. H. Freeman.

RELATIVITY AND ITS ROOTS

Banesh Hoffmann

In RELATIVITY AND ITS ROOTS, the Einstein scholar Banesh Hoffmann tells of the successive insights that led to both the special and general theories of relativity and explains those theories for the layperson.
From the geometrical and cosmological ideas of the Greeks to later advances by Galileo and Newton, we are taken on a journey that is both compelling and illuminating. Along the way, we encounter Kepler's laws, Galilean and Newtonian mechanics, and the work on electricity and magnetism by Faraday and Maxwell.

1983, 184 pages, 136 illustrations.

GRAVITATION

Charles W. Misner Kip S. Thorne John Archibald Wheeler

The first book to use the modern formalism and notation of differential geometry, GRAVITATION is also the first to document the revolutionary techniques developed during the past decade to test the theory of general relativity. Here is a work that is literally without peer.

"A pedagogic masterpiece."—Science

"This is one of the great books of science, a lamp to illuminate this Aladdin's cave of theoretical physics whose genie was Albert Einstein."—Science Progess

1973, 1,279 pages, 137 illustrations.

SELECTED PAPERS, 1945-1980, WITH COMMENTARY

Chen Ning Yang

A remarkable personal and professional chronicle by one of today's leading physicists, this is a collection of Chen Ning Yang's personally selected papers, supplemented by his insightful commentaries. Among the papers are his seminal work with T. D. Lee on the nonconservation of parity, for which they won the Nobel Prize.

"The book is a pleasure to read from start to finish, thanks to Yang's taste, style, and aesthetic sensitivity, coupled with his warmth and common touch."—Science

1983, 596 pages, 33 illustrations.

THE LIGHTER SIDE OF GRAVITY

Jayant V. Narlikar

As simple as walking on earth and as complicated as warped spacetime, the mysterious force of gravity holds the universe together. In this popular account, the author captures the wondrous power of gravity in humorous anecdotes, illustrations, and exciting observations from astronomy.

"If only there were more scientists who understood . . . the real principles of science, and if only there were more scientists who could write like Narlikar."—New Scientist

1982, 194 pages, 124 illustrations.

W. H. FREEMAN AND COMPANY, 4419 West 1980 South, Salt Lake City, Utah 84104

RELATIVITY AND ITS ROOTS
☐ ISBN 0-7167-1478-7 $16.95 (cloth)
☐ ISBN 0-7167-1510-4 $9.95 (paper)
SELECTED PAPERS, 1945-1980, WITH COMMENTARY
☐ ISBN 0-7167-1406-X $39.95 (cloth)
☐ ISBN 0-7167-1407-8 $19.95 (paper)
GRAVITATION
☐ ISBN 0-7167-0344-0 $43.50 (cloth)
THE LIGHTER SIDE OF GRAVITY
☐ ISBN 0-7167-1343-8 $17.95 (cloth)
☐ ISBN 0-7167-1344-6 $9.95 (paper)

☐ I enclose a check made payable to W. H. Freeman and Company. (Please add $1.50 for shipping and handling; New York, California, and Utah residents add appropriate sales tax.)
Charge my ☐ MasterCard ☐ VISA Exp. Date _____ _____
(All credit card orders must be signed.)

Signature _____

Name _____

Address _____

City _____ State _____ Zip _____

(0104)

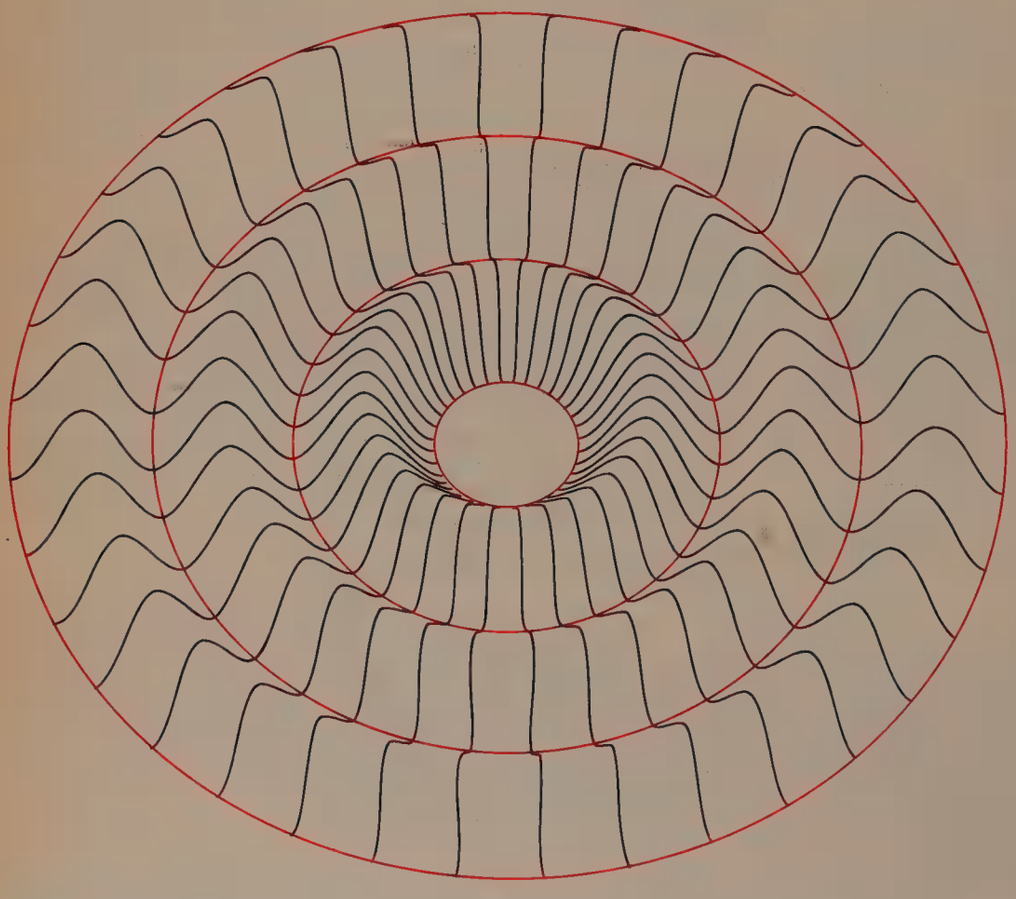
The Interpretation of Visual Illusions

The visual system apparently organizes ambiguous retinal images according to rules of inference that exploit certain regularities in the external world

by Donald D. Hoffman

Vision is a process of inference. What you see when you look around depends not only on what is there to be seen but also on how your visual system organizes and interprets the images that fall on your retinas. An intriguing demonstration of this aspect of perception is presented by the apparent surface that is formed by rotating a cosine wave around a vertical axis viewed obliquely [*see illustration on this page*]. When you first look at the figure, it appears to be organized into a set of raised concentric rings, with the boundaries between the rings delineated approximately by the colored circular contours. If you turn the page upside down, however, the organization changes: now each colored contour, instead of lying in a trough between two rings, appears to trace the crest of a ring. (Try it.) Evidently the visual system does more than passively transmit signals to the brain. It actively takes part in organizing and interpreting them.

This finding raises three questions. First, why does the visual system need to organize and interpret the images formed on the retinas? Second, how does it remain true to the real world in the process? Third, what rules of inference does it follow? The answers to these questions call for a closer examination of such figures.



AMBIGUOUS SURFACE is made by rotating a cosine wave about a vertical axis. The surface initially appears to be organized into raised concentric rings, with the colored circular contours lying in the troughs between the rings. When the page is turned upside down, however, the organization appears to change: each colored contour is now seen to trace the crest of a ring.

One reason the visual system organizes and interprets retinal images is simply that many possible configurations in the real world are consistent with any given retinal image. In other words, retinal images need organization and interpretation because they are fundamentally ambiguous. Their ambiguity is due in part to the fact that the world is three-dimensional and each retina is essentially two-dimensional. To describe the world in its full three-dimensional glory necessarily involves some rather sophisticated inferences by the visual system, inferences that for the most part proceed without any conscious awareness. For example, the cosine surface at the right, like your retinal image of it, is two-dimensional. Yet it appears, quite compellingly, as three-dimensional. The appearance of depth is entirely inferred, or, to put it another way, hallucinated. This conclusion should be cause for some concern. If, as I suggest, such hallucinations are not an exception but the rule, and if they are in fact a necessary concomitant of visual perception, how can one justify one's faith in perception? How is it still possible that in general seeing is believing?

What is needed for an understanding of vision, therefore, is an explanation of why such visual inferences usually bear a nonarbitrary relation to the real world. A promising line of investigation begins with the observation that the visible world, far from being completely chaotic, obeys certain laws and exhibits numerous regularities. If the visual system is adapted to exploit these laws and regularities in its organization and interpretation of retinal images, and if it is constrained somehow to prefer the interpretation that is most credible, given both the image and a knowledge of these laws and regularities, then it might be possible to understand how it is that one's visual hallucinations bear a nonarbitrary and even useful relation to the external world.

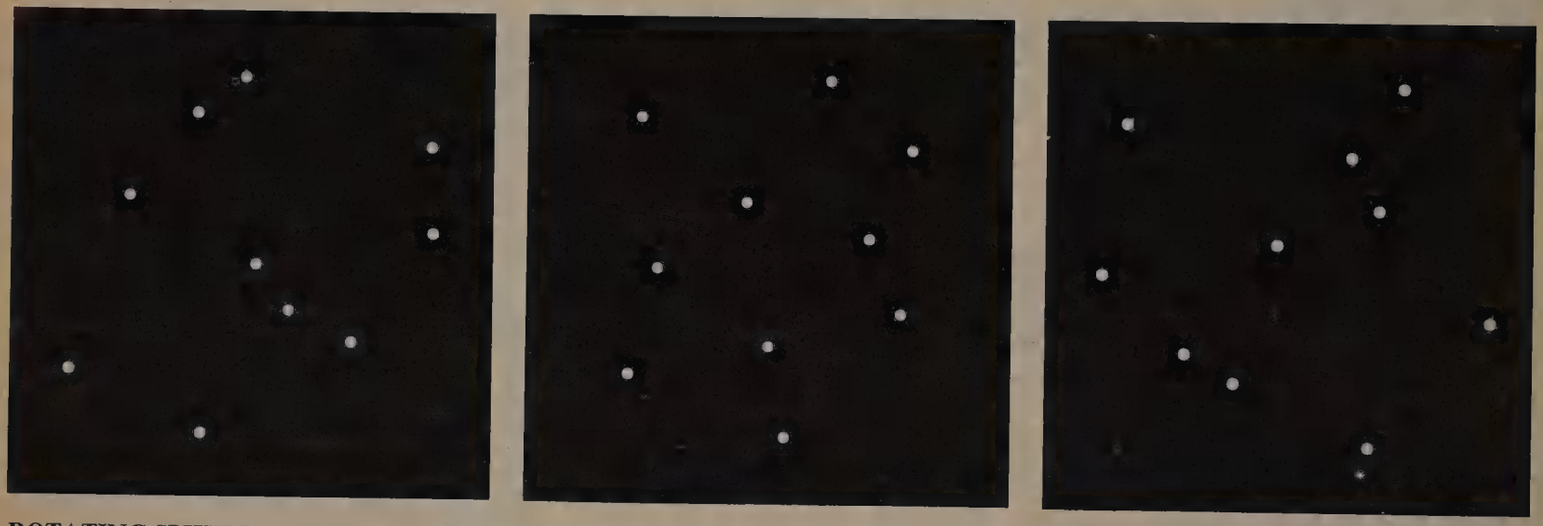
A particularly clear example of this approach is the research into visual mo-


100% GRAIN NEUTRAL SPIRITS. 94.6 PROOF. IMPORTED BY SOMERSET IMPORTERS, LTD., N.Y. 1983
IMPORTED
Tanqueray
SPECIAL DRY
Suddenly, the idea of importing a Christmas tree from England begins to make sense.
Tanqueray Gin. A singular experience.
Send a gift of Tanqueray Gin anywhere in the U.S.A. Call 1-800-528-6148. Void where prohibited.

tion perception done by Shimon Ullman of the Massachusetts Institute of Technology. Ullman has explored the remarkable ability of the human visual system to perceive the correct three-dimensional structure and motion of an object solely from its moving two-dimensional projection, an ability Hans Wallach and Donald N. O'Connell of Swarthmore College call the kinetic-depth effect. For instance, if a transparent beach ball with tiny light bulbs mounted randomly on its surface is set spinning in a dark room, one immediately perceives the correct spherical layout of the lights [*see upper illustration below*]. When the spinning stops, so does the perception of the spherical array. How does one see the correct three-dimensional structure when infinitely many three-dimensional structures are consistent with the moving two-dimensional retinal projection? Ullman showed mathematically that if the visual system exploits the laws of projection, and if it exploits the fact that the world contains rigid objects, then in principle a unique and correct interpretation can be obtained. In particular he showed that three views of four noncoplanar light bulbs are enough to solve the problem. The key point is that an inference rule, based on a law (the law of projection) and a regularity (namely the fact that the world includes rigid objects), enables the visual system to make a correct interpretation.

At this stage, however, a puzzle arises. The same mathematical precision that shows the rigidity regularity is sufficient in principle to interpret the rotating beach ball also shows the rigidity regularity by itself is insufficient to interpret a similar display. This display was first devised by Gunnar Johansson of the University of Uppsala as an example of what he calls biological motion [see "Visual Motion Perception," by Gunnar Johansson; SCIENTIFIC AMERICAN, June, 1975]. Johansson put small light bulbs on the major joints of a person and took motion pictures as the person moved about in a dark room. A single frame of such a film looks like a random collection of white dots on a black background. When the film is set in motion, however, one immediately sees the correct three-dimensional structure of the dots and recognizes that there is an invisible person walking about [*see lower illustration below*].

When my colleague Bruce E. Flinchbaugh, who is now at Bell Laboratories, and I considered this problem, what puzzled us was that it is possible to see the correct three-dimensional structure even though, according to Ullman's results, one lacks the appropriate information to do so. To infer a correct three-dimensional structure on the basis of the rigidity regularity it is necessary to have three snapshots of at least four nonco-



ROTATING SPHERE is seen when the three dot patterns represented here are shown in rapid succession. The visual system seems to be adopting the most rigid three-dimensional interpretation for the moving dots that is consistent with the two-dimensional projections.



WALKING PERSON is seen when these dot patterns are shown in rapid succession. In this case the visual system seems to adopt the most rigid and planar three-dimensional interpretation that is consistent with the two-dimensional motions of the dots. The display is based on an experiment conducted by Gunnar Johansson of the University of Uppsala in which small light bulbs were put on a person's major joints (shoulder, elbow, wrist, hip, knee and ankle) and a motion picture was made as the person moved about in a dark room.

planar points in a rigid configuration. In biological motion displays, on the other hand, at best only pairs of points are rigidly connected, such as the ankle and the knee or the knee and the hip. Rigid quadruplets of points just do not exist.

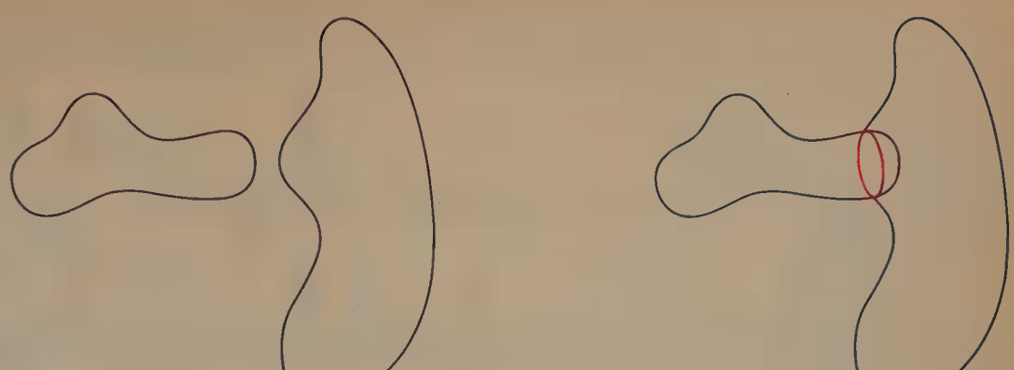
The rigidity regularity, then, is insufficient by itself, leading us to ask: What further regularity might the visual system be exploiting? After several false starts it occurred to us there is an anatomical regularity that might do the trick. Each weight-bearing limb of most animals is constrained, because of the construction of its joints, to swing in a single plane in a normal gait. We call this the planarity regularity.

In fact, the planarity regularity is sufficient to correctly interpret biological-motion displays of gait. The correct three-dimensional structure can be inferred either from three snapshots of two points swinging rigidly in a plane or from two snapshots of three points (such as an ankle, a knee and a hip) forming rigid pairs and swinging in one plane. These results comport nicely with Johansson's observation that only two or three frames of his films need be seen before subjects correctly perceive the biological motion. In addition it turns out not only that all three-dimensional motions governed by the planarity regularity can be given a correct interpretation but also that whenever an interpretation is found for image motion based on the planarity regularity or the rigidity regularity the interpretation is correct.

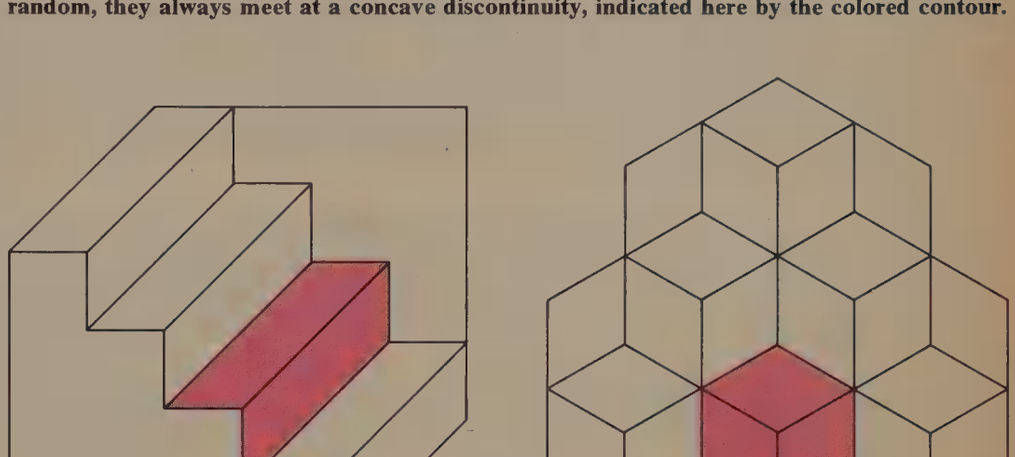
In short, the probability that the interpretation is wrong is zero, assuming infinite resolution in the image, or slightly greater than zero given less than perfect resolution. Hence nonrigid structures cannot masquerade as rigid ones, and nonplanar motions cannot be misconstrued as planar ones. Once again laws and regularities prove to be central in explaining how the visual system achieves a unique and correct interpretation of a retinal image.

Let us now return to the cosine surface. Its main interest is that it reveals the visual system organizing shapes into parts, an organization that is quite useful for the task of recognizing an object from its shape. The cosine surface also reveals that turning a shape upside down can alter this organization. Is the visual system, then, capricious in its organization? That is unlikely. If it is not governed by whim, however, it must be governed by rules for defining parts. And if the rules are not to be arbitrary, they must be grounded in some law or regularity in the external world.

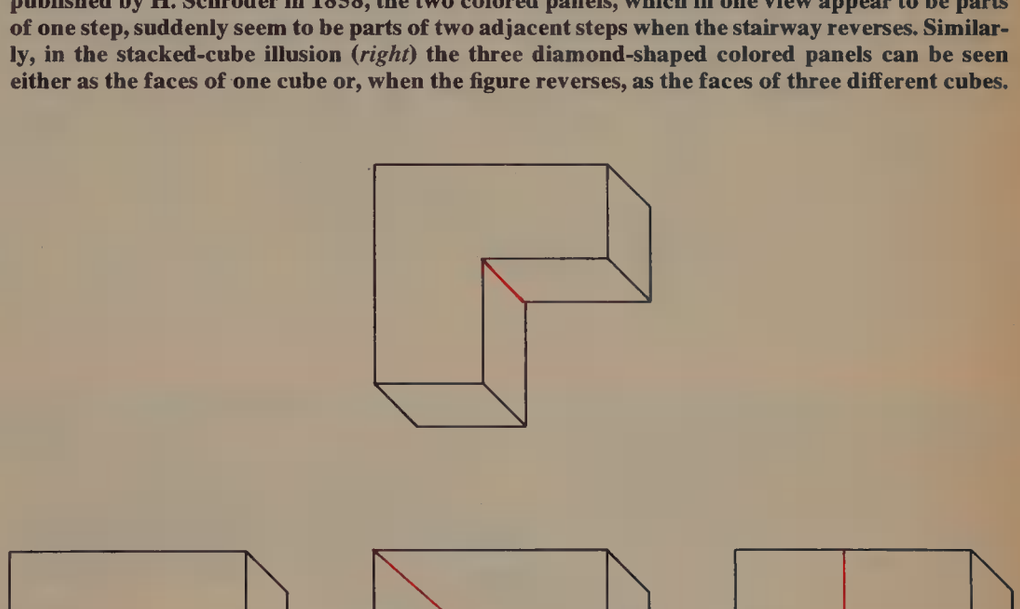
This line of reasoning led Whitman A. Richards of M.I.T. and me to seek a law or regularity that could motivate a set of rules for partitioning surfaces. The regularity we found to be relevant is the following transversality regularity:



TRANSVERSALITY, a kind of regularity commonly observed in the external world, underlies a unified account of several visual illusions. According to the rule of transversality (as defined by Whitman A. Richards and the author), when any two surfaces penetrate each other at random, they always meet at a concave discontinuity, indicated here by the colored contour.



PARTITIONING RULE based on the transversality regularity is demonstrated with the aid of these two figures that reverse when they are looked at steadily. In both cases the apparent boundaries of the different parts of the perceived shape change when the "figure" becomes the "ground" and vice versa. For example, in the case of the reversing-stairway illusion (*left*), first published by H. Schröder in 1858, the two colored panels, which in one view appear to be parts of one step, suddenly seem to be parts of two adjacent steps when the stairway reverses. Similarly, in the stacked-cube illusion (*right*) the three diamond-shaped colored panels can be seen either as the faces of one cube or, when the figure reverses, as the faces of three different cubes.



ELBOW-SHAPED BLOCKS show that the rule partitioning shapes at concave discontinuities is appropriately conservative. The rule does not give a closed contour on the top block because three different perceived partitions seem possible, as illustrated by the bottom three blocks.

SIGN WITH A WINNING CLUB.

FORD CLUB WAGONS

BEST-BUILT AMERICAN TRUCKS

Ford leads off with the best-built van/passenger van. With out-front design, Ford Club Wagons can out-tow and out-carry any station wagon. There's passenger seating for up to 12 (15 in the roomy Super Wagon)! It's built tough, with body-on-frame construction. No wonder Ford is a winning club!

Get it together–Buckle up.



TOWS UP TO 3,000 LB. MORE THAN CHEVY OR DODGE!
Ford scores with power, rated up to 10,000 lb. towing capacity when properly equipped. Plus the most powerful gas or diesel engines.

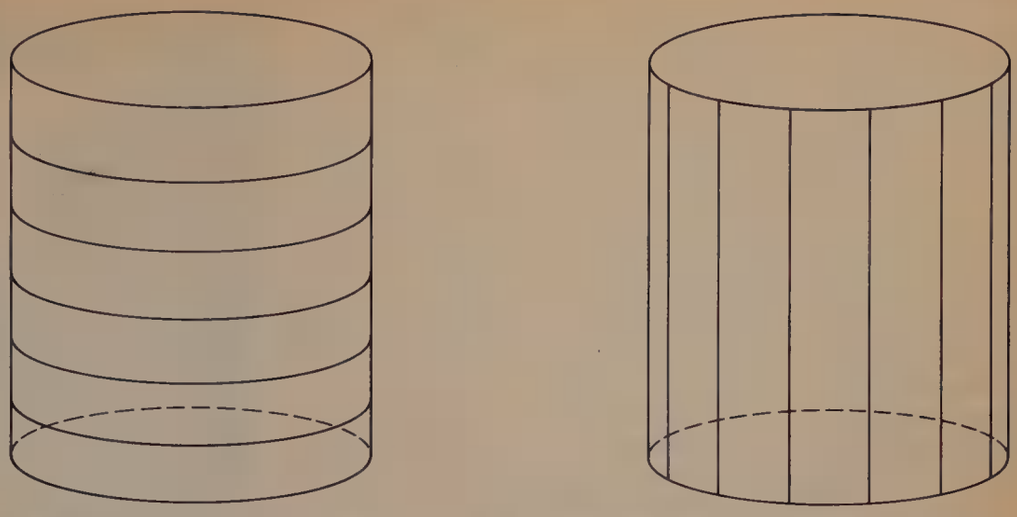
Quality is Job 1.

Ford's commitment to quality results in the best-built American trucks, based on a survey of owner problems during the first three months of ownership of 1983 trucks. And the commitment continues in 1984.

BUILT FORD TOUGH

Ford

When two arbitrarily shaped surfaces are made to penetrate each other at random, they always meet at a contour of concave discontinuity of their tangent planes [*see top illustration on page 157*]. Although the transversality regularity may sound esoteric, it is actually a familiar part of everyday experience. A straw in a soft drink, for instance, forms a circular concave discontinuity where it meets the surface of the drink. A candle in a birthday cake, the tines of a fork in a piece of steak, a cigarette in a mouth—all are examples of this ubiquitous regularity.



LINES OF CURVATURE are easily pictured on an idealized cylindrical drinking glass. The lines of greatest curvature (*left*) are circles; the lines of least curvature (*right*) are straight lines.

On the basis of the transversality regularity one can propose a first rule for partitioning a surface: Divide a surface into parts along all contours of concave discontinuity. This rule cannot help with the cosine surface because it is entirely smooth. The rule must first be generalized somewhat, as will be done below. In its nongeneralized form, however, it can elucidate several well-known perceptual demonstrations.

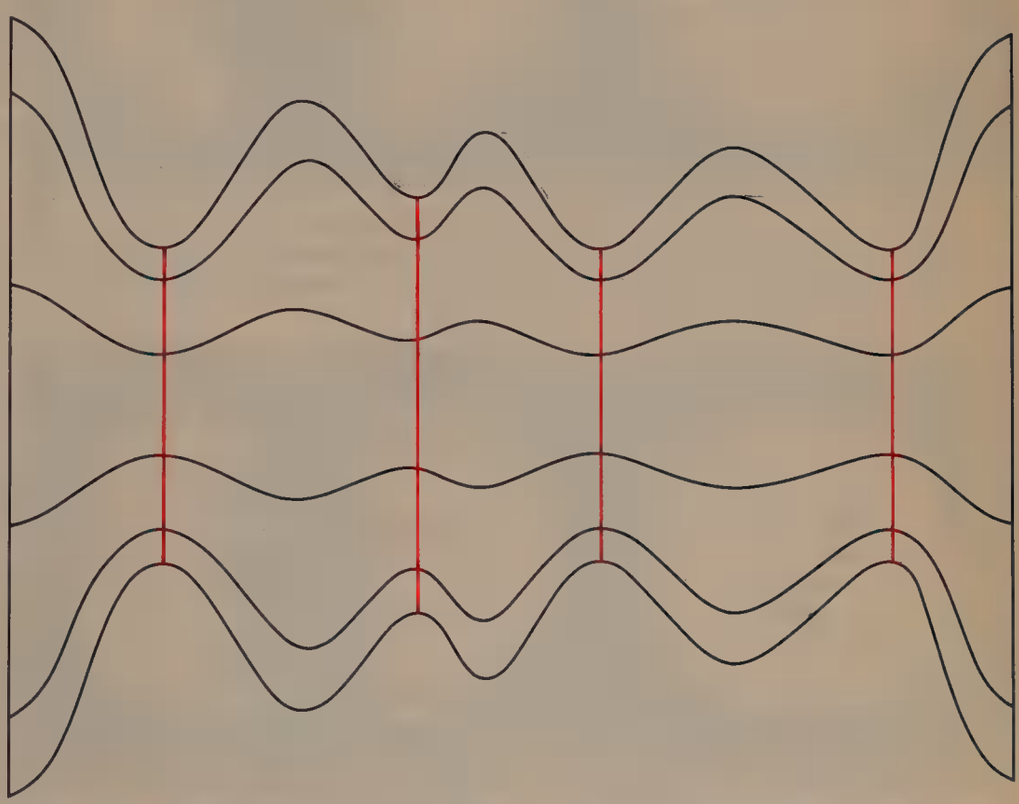
For example, the rule makes the obvious prediction that the parts of the staircase shown in the middle illustration on page 157 are its steps, each step lying between two successive lines of concave discontinuity in the staircase. The rule also makes a less obvious prediction. If the staircase undergoes a perceptual reversal, such that the "figure" side becomes "ground" and vice versa, then the step boundaries must change. This conclusion follows because only concave discontinuities define the step boundaries, and what looks like a concavity from one side of a surface must look like a convexity from the other. Thus when the staircase reverses, convex and concave discontinuities must reverse roles, leading to new step boundaries. You can test this prediction yourself by looking at the step that has color on each of its two faces. When the staircase appears to reverse, note that the colored panels are no longer on a single step but rather on adjacent steps.

This prediction can be confirmed with a more complicated demonstration such as the stacked-cubes test seen in the same illustration. The three colored faces, which at first appear to be on one cube, are seen to be on three cubes when the figure reverses.

A further prediction follows from this simple partitioning rule. If the rule does not define a unique partition of some surface, then the appropriate way to divide the surface into parts should be perceptually ambiguous (unless there are additional rules that can eliminate the ambiguity). A clear confirmation of this prediction can be seen with reference to the elbow-shaped block in the bottom illustration on page 157. The only concave discontinuity is the vertical line in the crook of the elbow. As a consequence the rule does not define a unique partition of the block. Perceptually there are three plausible ways to cut the block into parts. All three ways rely on the contour defined by the partitioning rule, but they complete it along different paths.

Even this simple partitioning rule leads to interesting insights into the perception of shape. To explore the cosine surface and other smooth surfaces, however, the rule must be generalized. This requires a brief digression into the differential geometry of surfaces in order to understand three important concepts: surface normal, principal curvature and line of curvature. Fortunately, although these concepts are quite technical, they can readily be given an intuitive characterization.

The surface normal at a point on a surface can be thought of as a needle of unit length sticking straight out of the surface at that point, much like the spines on a sea urchin. All the surface normals at all points on a surface are collectively called a field of surface normals. Usually there are two possible fields of surface normals on a surface; they can be either outward-pointing or inward-pointing. For example, a sphere can have the surface normals all pointing radially out like spines or all pointing in toward its center. Let us adopt the convention that the field of surface normals is always chosen to point into the figure. Thus a baseball has inward normals whereas a bubble under water has outward normals. Reversing the choice of figure and ground on a surface im-



PART BOUNDARIES, as defined by the generalized, smooth-surface partitioning rule, are represented by the colored contours on this arbitrarily shaped surface. The black lines are the lines of greatest curvature whose minimums give rise to the colored partitioning contours.

plies a concomitant change in the surface normals. A reversal of the field of surface normals induces a change in sign of each of the principal curvatures at every point on the surface.

It is often important to know not only the surface normal at a point but also how the surface is curving at the point. The 18th-century Swiss mathematician Leonhard Euler discovered that at any point on any surface there is always a direction in which the surface curves the least and a second direction, always at right angles to the first, in which the surface curves the most. (In the case of a plane or a sphere the surface curvature is identical in all directions at every point.) These two directions are called principal directions, and the corresponding surface curvatures are called principal curvatures. By starting at some point and always moving in the direction of the greatest principal curvature one traces out a line of greatest curvature. By moving instead in the direction of the least principal curvature one traces out a line of least curvature. On a drinking glass the family of lines of greatest curvature is a set of circles around the glass. The lines of least curvature are straight lines running the length of the glass [*see top illustration on preceding page*].

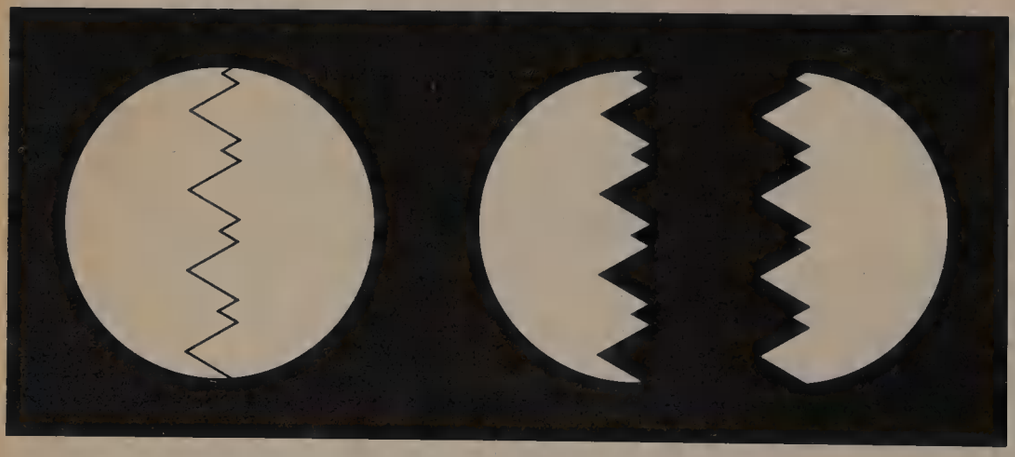
With these concepts in mind the transversality regularity extends easily to smooth surfaces. Suppose wherever a surface has a concave discontinuity one smoothes the discontinuity somewhat, perhaps by stretching a taut skin over it. Then a concave discontinuity becomes, roughly speaking, a contour where the surface has locally the greatest negative curvature. More precisely, the generalized version of transversality suggests the following generalized partitioning rule for surfaces: Divide a surface into parts at negative minimums of each principal curvature along its associated family of lines of curvature [*see bottom illustration on preceding page*].

This rule partitions the cosine surface along the colored circular contours. It also explains why the parts are different when the page is turned upside down: the visual system then reverses its assignment of figure and ground on the surface (perhaps owing to a preference for an interpretation that places the object below the observer's viewpoint rather than above it). When figure and ground reverse, so does the field of surface normals, in accordance with the convention mentioned above. Simple calculations show, however, that when the normals reverse, so does the sign of the principal curvatures. As a result minimums of the principal curvatures must become maximums and vice versa. Since minimums of the principal curvatures are used for part boundaries, it follows that these part boundaries must also move. In sum, parts appear to change because the partitioning rule, motivated by the transversality regularity, uses minimums of the principal curvatures, and because these minimums relocate on the surface when figure and ground reverse.

The transversality regularity, in short, provides an underlying unity for explanatory accounts of the perception of parts in both smooth and rough surfaces. It also underlies an explanation of another well-known class of visual illusions: reversing plane curves. A good example of this phenomenon is the reversing figure devised by Fred Attneave of the University of Oregon [*see upper illustration on this page*]. He found that by simply scribbling a line through a circle and separating the two halves one can create two very different-looking contours. Evidently, as Attneave points out, the appearance of the contour depends on which side is taken to be part of the figure, not on any prior familiarity with the contour [see "Multistability in Perception," by Fred Attneave; SCIENTIFIC AMERICAN, December, 1971].



REVERSING PLANE CURVE, constructed by Fred Attneave of the University of Oregon by scribbling a line through a circle and separating the two halves, shows that the apparent shape of the resulting contour depends on which side of the line is perceived as the figure.

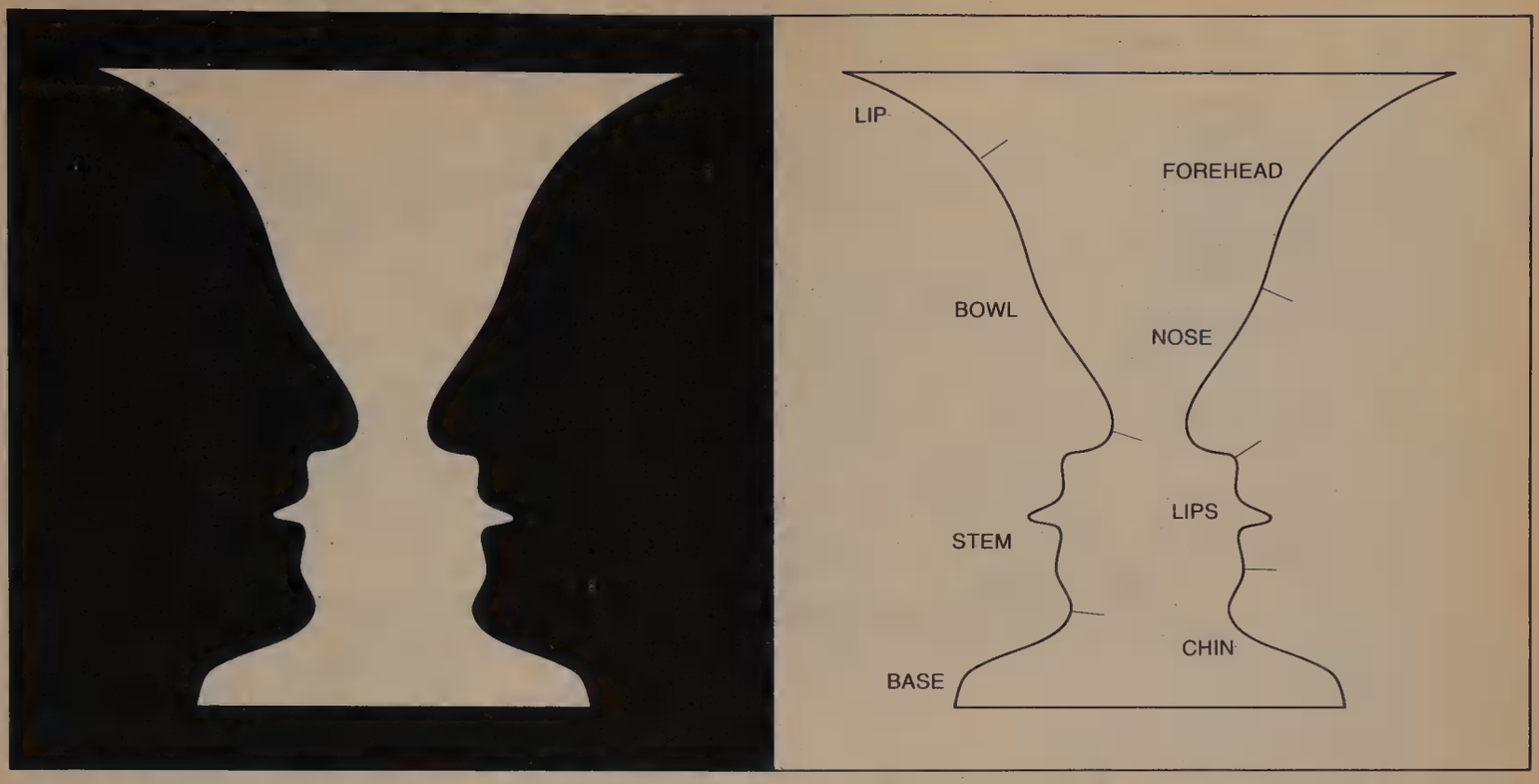


SIMILAR REVERSING FIGURE can be made with a plane curve that is not smooth. One can see the resulting jagged contour either as an alternating chain of tall and short mountains or, in the reversed figure-ground assignment, as a chain of tall mountains with twin peaks.

How does the transversality regularity explain this phenomenon? The answer involves three steps: (1) a projection of the transversality regularity from three dimensions onto two dimensions, (2) a brief digression on the differential geometry of plane curves and (3) the formulation of a partitioning rule for plane curves.

The two-dimensional version of the transversality regularity is similar to the three-dimensional version. If two arbitrarily shaped surfaces are made to penetrate each other at random, then in any two-dimensional projection of their composite surface they will always meet in concave cusps. To paraphrase it loosely, concave cusps are always formed in a silhouette at points where one part stops and another begins. This suggests the following partitioning rule for plane curves: Divide a plane curve into parts at concave cusps. This rule cannot apply to Attneave's demonstration because his demonstration relies on a contour that is everywhere smooth. The rule must again be generalized. Nevertheless, in its nongeneralized form it can account for a version of Attneave's demonstration that is not everywhere smooth.

In the lower illustration at the left the same jagged contour can look either like an alternating chain of tall and short mountains or, for the reversed



FACE-GOBLET ILLUSION, devised by Edgar Rubin in about 1915, can be seen either as a pair of facial profiles or as a goblet (*left*). If a face is taken to be the figure, partitioning the figure by reference to minimums of curvature divides the contour into chunks corresponding to a forehead, a nose, a pair of lips and a chin; if the goblet is taken to be the figure, defining the part boundaries by minimums of curvature divides the contour into a lip, a bowl, a stem and a base (*right*). Principal-normal lines in both cases point into the figure.

figure-ground assignment, like a chain of tall mountains with twin peaks. The contour is carved into parts differently when figure and ground reverse because the partitioning rule uses only concave cusps for part boundaries. What is a concave cusp if one side of the contour is figure must become a convex cusp when the other side is figure, and vice versa. There is a parallel between this example and the reversible staircase discussed above.

Before generalizing the rule to smooth contours let us briefly review two concepts from the differential geometry of plane curves: principal normal and curvature. The principal normal at a point on a curve can be thought of as a unit-length needle sticking straight out of the curve at that point. All the principal normals at all points on a curve together form a field of principal normals. Usually there are two possible fields of principal normals, one field on each side of the curve. Let us adopt the convention that the field of principal normals is always chosen to point into the figure side of the curve. Reversing the choice of figure and ground on a curve implies a concomitant change in the field of principal normals. What is important to note is that because of the convention forcing the principal normals to point into the figure, concave parts of a smooth curve have negative curvature and convex parts have positive curvature.

It is an easy matter now to generalize the partitioning rule for plane curves. Suppose wherever a curve has a concave cusp one smoothes the curve a bit. Then a concave cusp becomes a point of negative curvature having, locally, the greatest absolute value of curvature. This observation leads to the following generalized partitioning rule: Divide a plane curve into parts at negative minimums of curvature.

Now it is possible to explain why the two halves of Attneave's circle look so different. When figure and ground reverse, the field of principal normals also reverses in accordance with the convention, and when the principal normals reverse, the curvature at every point on the curve must change sign. In particular, minimums of curvature must become maximums and vice versa. This repositioning of the minimums of curvature leads to a new partitioning of the curve by the partitioning rule. In short, the curve looks different because it is organized into fundamentally different chunks, or units. Note that if one chooses to define part boundaries by inflections, or by both maximums and minimums of curvature, then the chunks would not change when figure and ground reverse.

A clear example of two very different partitions for one curve can be seen in the famous face-goblet illusion devised by Edgar Rubin in about 1915 [*see illustration above*]. If a face is taken to be figure, the minimums of curvature divide the curve into chunks corresponding to a forehead, nose, upper lip, lower lip and chin. If instead the goblet is taken to be figure, the minimums are repositioned, dividing the curve into new chunks corresponding to a base, a couple of parts of the stem, a bowl and a lip on the bowl. It is probably no accident that the parts defined by minimums are often easily assigned verbal labels.

Demonstrations have been devised that, like the face-goblet illusion, allow more than one interpretation of a single contour but that do not involve a figure-ground reversal. A popular example is the rabbit-duck illusion [*see top illustration on next page*]. Because such illusions do not involve a figure-ground reversal, and because as a result the minimums of curvature never change position, the partitioning rule must predict that the part boundaries are identical for both interpretations of each of the contours. This prediction is easily confirmed. What is an ear on the rabbit, say, becomes part of a bill on the duck.

If the minimums rule for partitioning curves is really obeyed by the human visual system, one would expect it to predict some judgments of shape similarity. One case in which its prediction is counterintuitive can be seen in the bottom illustration on the next page. Look briefly at the single half-moon on the left side of the illustration. Then look quickly at the two half-moons at the right and decide which one seems more like the first one. In an experiment done on several similar figures, Aaron F. Bobick of M.I.T. and I found that almost all subjects chose the half-moon at the lower right as the more similar one. Yet if you look again, you will find that the bounding contour for the half-moon at

Aha...



GAMES FOR THINKERS

An exciting, new way to learn creative problem solving! Games designed by university professors improve thinking skills using fun, strategy and challenging competition. Like chess, each game can be played at many levels from young children to intelligent adults. Fascinating for everyone!

Write for free catalog and studies that show how WFF 'N PROOF Games can:

- double math achievement
- cut school absenteeism by ⅔ and
- raise I.Q. scores by 20 points

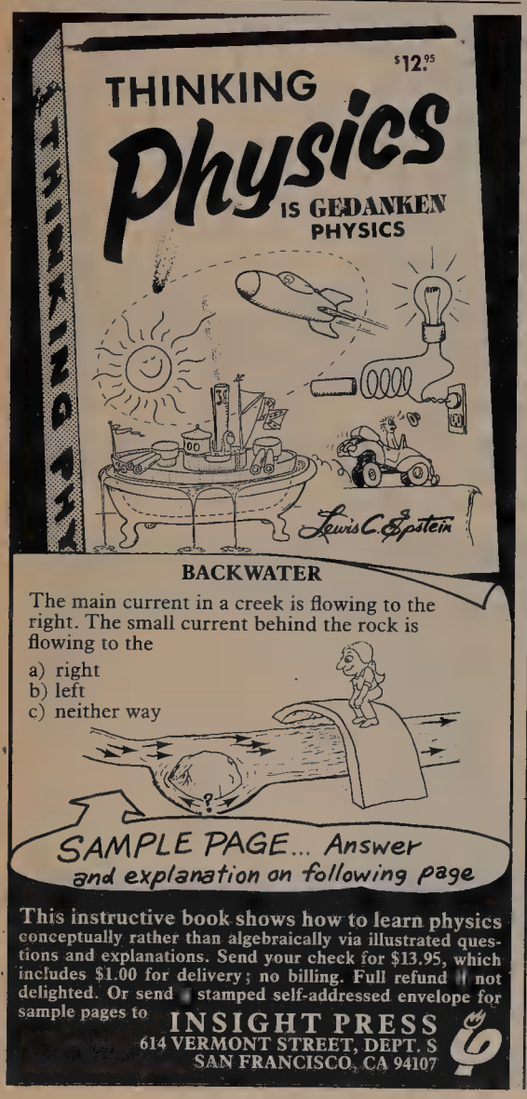
ORDER YOUR GAMES FOR THINKERS TODAY!

WFF'N PROOF (logic)	$16.00
QUERIES 'N THEORIES (sci. method)	16.00
EQUATIONS (creative mathematics)	13.00
ON-SETS (set theory)	13.00
ON-WORDS (word structures)	13.00
PROPAGANDA (social studies)	13.00
CONFIGURATIONS (geometry)	7.75
Complete 7-game Special	**79.95**

All prices include postage and handling.
Satisfaction Guaranteed

Order from: **WFF 'N PROOF**

1490-FD South Blvd., Ann Arbor, MI 48104

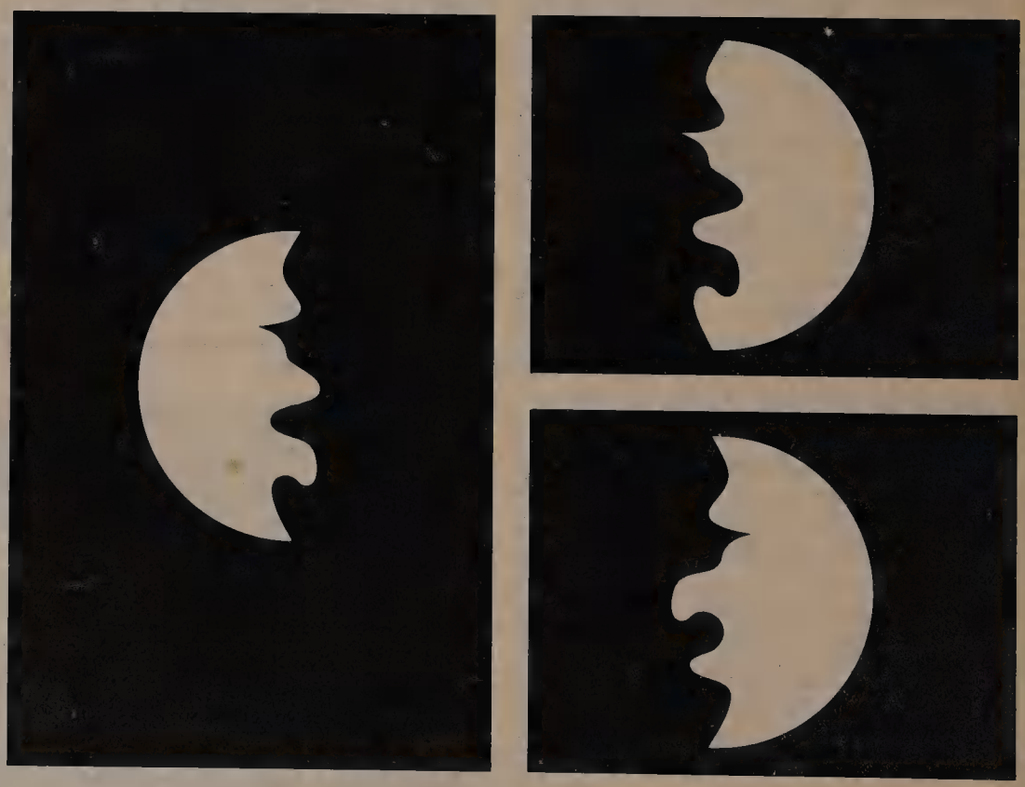




REVERSING-ANIMAL ILLUSION does not involve a reversal of figure and ground. Accordingly the part boundaries defined by the minimums of curvature do not change position when the interpretation changes. The rabbit's ears turn into the duck's bill without moving.

the upper right is identical with that of the left half-moon, only it is figure-ground reversed. The bounding contour of the lower half-moon has been mirror reversed, and two parts defined by minimums of curvature have been swapped. Why does the lower one still look more similar? The minimums rule gives a simple answer. The lower contour, which is not figure-ground reversed from the original contour, has the same part boundaries. The upper contour, which is figure-ground reversed from the original, has different part boundaries.

In summary, vision is an active process whose function is to infer useful descriptions of the world from changing patterns of light falling on the retinas. These descriptions are reliable only to the extent that the inferential processes building them exploit regularities in the visual world, such as rigidity, planarity and transversality. The discovery of relevant regularities and the mathematical investigation of their power in guiding visual inferences are promising directions for the investigator seeking an understanding of human vision.



HALF-MOON TEST demonstrates that judgments of the similarity of shapes can be correctly predicted by the minimums-of-curvature partitioning rule. At first glance the half-moon at the lower right seems to resemble the single half-moon at the left more than the one at the upper right does. Closer inspection, however, reveals that the bounding contour of the upper-right half-moon is identical with that of the half-moon at the left, whereas the bounding contour of the lower-right half-moon has been mirror reversed and has also had two parts interchanged.

Save $80 on Radio Shack's New TRS-80® Color Computer 2!



16K Standard BASIC
Reg. 239.95
159.95
26-3026

16K Extended BASIC
Reg. 319.95
239.95
26-3027

Use Your CitiLine Card

As Low As $25 Per Month On CitiLine Credit

A Smart Investment. Give your family a colorful Christmas this year with our TRS-80 Color Computer 2. It's the perfect way to join the computer revolution. And, at $80 off, it's the affordable way, too!

Easy to Use. Just connect your TV, plug in a ready-to-run Program Pak™ and you're all set. Prepare a household budget, monitor your investments, teach your kids or play thrilling outer space and sports games.

Deluxe Features. Both models feature a typewriter-quality keyboard in an attractive white case. Use the built-in Color BASIC language and our entertaining 308-page manual to start writing your own programs with color and sound. Want to create stunning high-resolution graphics and write advanced programs? Then select our sale-priced 16K Extended BASIC Color Computer 2. Create incredible drawings and designs, business and engineering charts, even animation with simple one-line commands.

Expandable. Best of all, Color Computer 2 can expand easily as your computing skills grow. Add a cassette recorder to store and load your own programs, a telephone modem to access tomorrow's news and stock reports, a printer, joysticks or a color mouse for more exciting game action, disk drives and more.

The Perfect Gift to Give . . . Or Get! Come into a Radio Shack Computer Center, participating store or dealer near you and get a demonstration of our spectacular Color Computer 2.

Radio Shack®

The biggest name in little computers®

A DIVISION OF TANDY CORPORATION

Send for your free 1984 TRS-80 Computer Catalog.

Mail To: Radio Shack, Dept. 84-A-702
300 One Tandy Center, Fort Worth, Texas 76102

NAME _____

COMPANY _____

ADDRESS _____

CITY _____ STATE _____ ZIP _____

TELEPHONE _____

Sale ends Dec. 31, 1983. TV not included. Prices apply at participating Radio Shack stores and dealers.

THE AMATEUR SCIENTIST

Funny things happen when drops of oil or other substances are placed on water

by Jearl Walker

What happens when a drop of oil or some other liquid is placed on the surface of water? If it is soluble in water, it disappears quickly into the bulk liquid. If it is insoluble, as paraffin oil is, it remains where it is placed. If it has an intermediate solubility, as olive oil has, it may spread over the surface of the water, eventually becoming invisibly thin; it will pool or bead on the surface if more drops are added. What determines the behavior of the drop?

The answer to the question remains somewhat uncertain. Part of it is that the behavior of a drop placed on a water surface is determined by the complex interactions (mostly electrical) of the molecules at the interface. Recently Frode Wissing of the Royal Dental College in Denmark sent me a manuscript about his experiments on the solubility of oleic acid (a major component of olive oil) and paraffin oil in water. His work, which was designed as a classroom demonstration, reveals several curious pieces of the solubility puzzle.

The solubility of a chemical compound in water depends on the extent of the bonding between the molecules of the compound and those of the water. The degree of solubility results from a competition between the bonds that hold each molecule together and the alternative bonding opportunities offered by the other substance.

Organic compounds range widely in their solubility in water. Life on the earth would not exist without this variability. The insoluble organic compounds have component groups of atoms that form few bonds (and in some cases none) with water molecules. Such groups are said to be hydrophobic, as is a molecule that is insoluble because of such groups. The term is misleading in that it implies a repulsion between the molecule (or a group) and the water. The effect arises not from repulsion but from the fact the bonding is so weak that the cohesion of the water keeps the hydrophobic compound out.

Paraffin oil is an example of a hydrophobic compound. It consists entirely of hydrophobic groups of carbon and hydrogen atoms. Since hydrocarbon groups form few bonds or none with water molecules, a drop of paraffin oil stays intact when it is placed on water. The drop pushes down on the water surface because of its weight, but it cannot dissolve in the water or even spread out over the surface.

Many other organic molecules are partially soluble in water because at least some of their atomic groups bond with water. The more such groups a compound has, the more soluble it is. For example, the sugar glucose dissolves readily in water because it has six hydroxyl (OH) groups that bond well to water molecules.

Oleic acid is intermediate to paraffin and glucose. One end of the molecule has a COOH group that bonds well to water. Such a group is said to be hydrophilic. The rest of the molecule, which is a chain of hydrocarbon groups, is hydrophobic. This dual nature of the oleic acid molecule is responsible for its ability to spread over the water in a layer only one molecule thick.

Wissing set out to illustrate the differences in solubility of paraffin oil (strictly hydrophobic) and oleic acid (partly hydrophilic). He did his experiments in petri dishes that were nine centimeters in diameter and absolutely clean. (Any contamination, even from a fingerprint, deposits a monolayer on the water surface and may alter the results.) A dish was placed on an overhead projector and surrounded by dark paper to make the light pass only through the dish. The arrangement yielded a magnification of about 10 for the events in the dish, more than enough for a class to follow as Wissing added drops of organic fluids to water in the dish. I repeated his experiments in my kitchen satisfactorily even though I substituted pure olive oil for the oleic acid.

Wissing first demonstrated the insolubility of paraffin oil by placing a drop of it on distilled water. Since the oil is insoluble, it merely rested on the water in a lenslike shape. As more drops were put on the water some coalesced to form a larger lens, but none spread over the surface or dissolved.

Wissing then did three similar experiments with dishes containing other solutions: hydrochloric acid, sodium hydroxide and ammonium hydroxide, each in the dilute concentration of .1 molar. In each dish the oil formed a lens, neither dissolving in the solution nor spreading over its surface.

Repeating the demonstrations with a drop of oleic acid, Wissing got more dramatic results. He had dyed the oleic acid with a small amount of Sudan III Red so that he could see it better. When he put a drop of oleic acid on pure water, the water surface developed a shock wave, as if someone had thumped the dish. The drop spread quickly over the surface, leaving a film so thin that even the red dye did not show up. Some of the acid also formed into a large lenslike drop and several smaller ones.

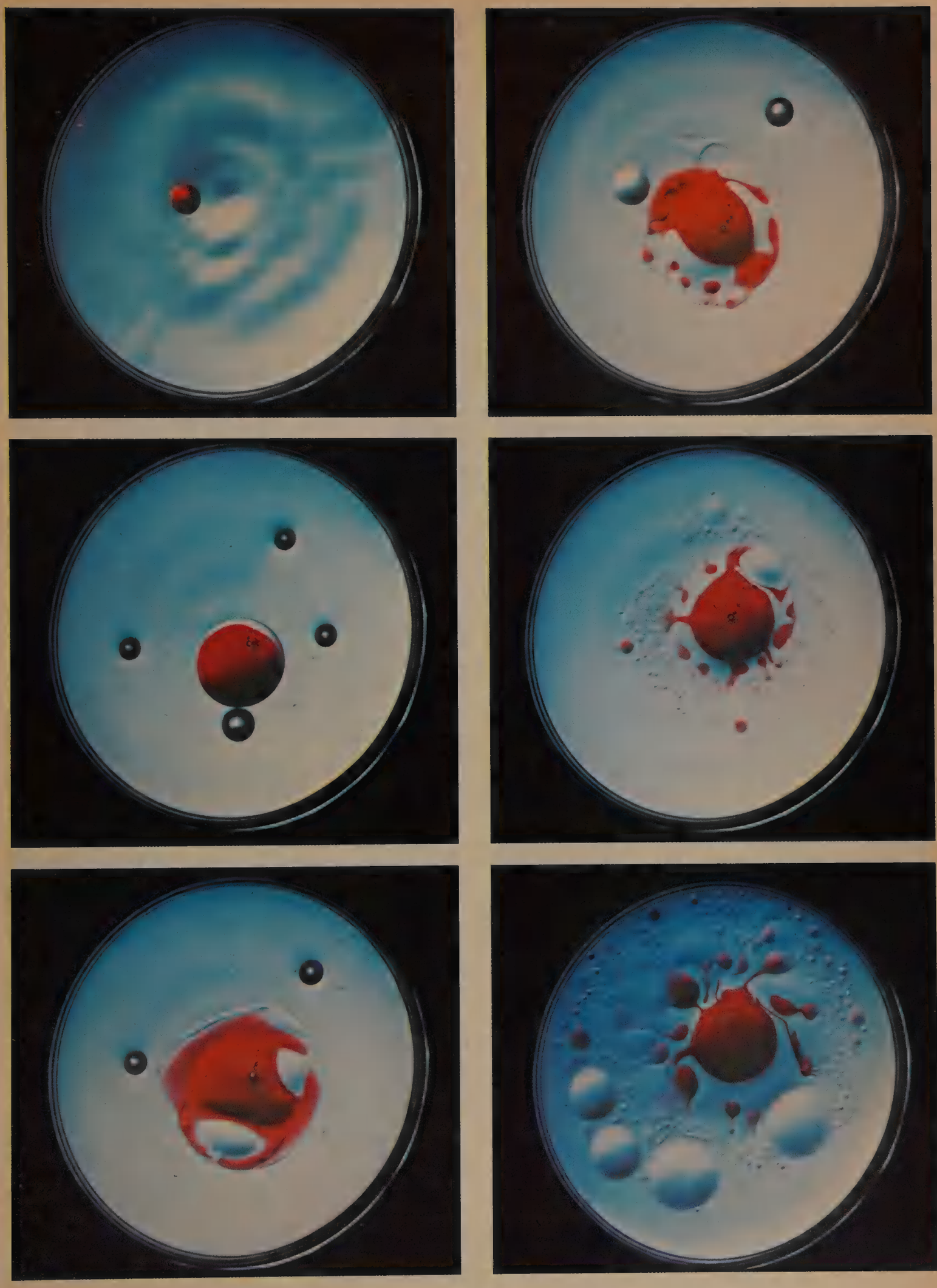
In my experiments I deployed a tiny drop of olive oil by flicking it from a straightened paper clip. (I had previously cleaned the clip with detergent and then heated one end to red heat.) The drop immediately disappeared as it spread over the surface, forming a layer much too thin to be observed directly or even to generate the kind of optical interference seen when an oil slick lies on water. The surface must have been almost completely covered with the olive oil, because a second drop held together for several minutes before spreading gradually into a layer that was quite visible in some places and at least thick enough in others for optical interference to create colored bands on it.

When Wissing added a drop of oleic acid to a dish of dilute hydrochloric acid, it again spread over the entire surface and formed a large drop. This time the lens was flatter, nearly filling the dish. Oleic acid added to a solution of sodium hydroxide formed an irregular lens. The acid flowed slowly outward from the drop in thin streams. After 20 minutes the lens had disappeared, leaving a turbid solution in the dish.

A drop added to the ammonium hydroxide zipped over the surface in small streams, each one bordered with clear zones, and then disappeared. The surface became calm after the acid stopped reacting with the ammonium hydroxide. Reaction products lay on the surface, but the bulk liquid showed no turbidity.

Wissing set himself the task of explaining those variations in terms of the forces acting between the oleic acid and the different solutions. The simplest force is gravity, which pulls a drop of oleic acid down into the solution, tending to spread it over the surface. Another force, arising from the displacement of the solution, is buoyancy.

More complex (and interesting) are



Frode Wissing's experiments in which a drop of oleic acid consumes paraffin drops on water

the many electrical interactions that can take place between the molecules in the drop and those in the solution. I shall discuss the simple quantum-mechanical models of these forces and then follow Wissing's application of the models to his experiments. Bear in mind, however, that all the models are flawed by the lack of detailed understanding of the quantum mechanics involved.

When a molecule has sites with a net electric charge, it is able to bond through electrostatic interactions. For example, a molecule that has a site of net positive charge attracts a molecule that has a site of net negative charge in what is termed an ionic bond. A second type of electrostatic attraction develops when one or both molecules have an electric moment. (This term describes the distribution of charge in a nonspherical molecule.) Although the molecule may be neutral, the center of its negative charge (from the electrons) does not coincide with the center of its positive charge (from the protons). Such a state sets up an electric field surrounding the molecule. When two such molecules sample each other's electric field, they are attracted.

A third possible attractive force arises when a molecule that has a strong electric moment is near another molecule that has no moment. The electric field from the first molecule shifts the charge distribution in the second molecule, creating an attraction between them. For example, the first molecule might shift the electron orbits in the second one so that the center of the negative charge is then farther away than the center of the positive charge.

A virtually ubiquitous interaction between molecules with permanent electric moments is the hydrogen bond. Here attention is directed to a hydrogen atom lying between the molecules and serving to hold them together. Water is cohesive because of the hydrogen bond. A water molecule consists of one oxygen atom and two hydrogen atoms, together forming a wide *V* with the oxygen at the apex. Although the molecule has no net charge, the electron associated with each hydrogen atom is strongly attracted to the large oxygen nucleus. The molecule thus has an internal charge distribution that creates an electric field around it.

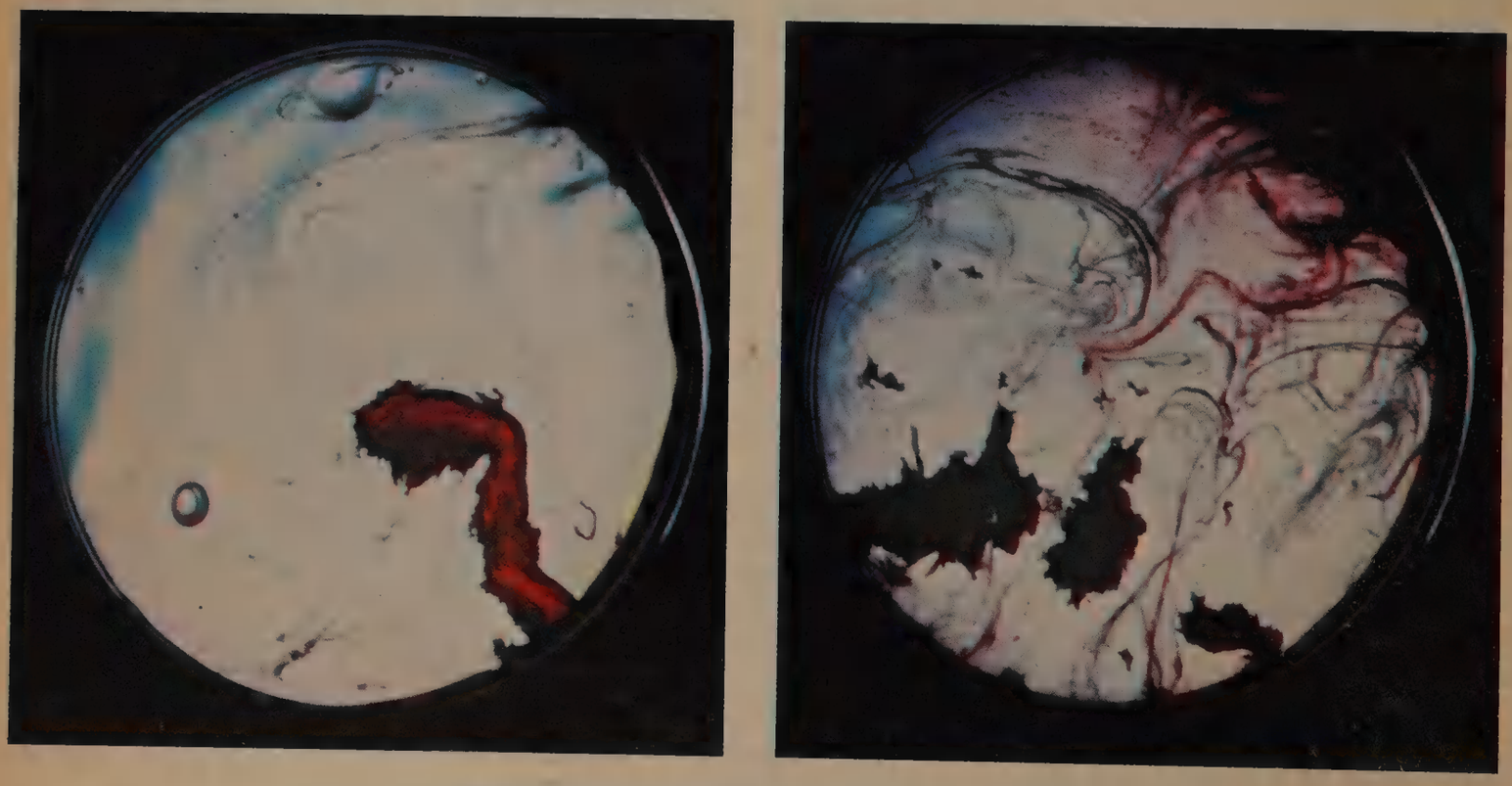
A simpler description is that the hydrogen ends of the *V* are left positive and the oxygen apex is left negative. One should resist the oversimplification of stating that the hydrogen atoms have been stripped of their electrons. The charge distribution is not that extreme. Often the charge distribution of a water molecule is said to be polar to indicate the presence of an electric moment.

Imagine two adjacent water molecules aligned as is shown in the top illustration on page 170. The molecule at the left presents one of its hydrogen atoms to the oxygen in the molecule at the right, thereby forming a hydrogen bond. The attraction is an electrical one in which the positive end of one molecule is pulled toward the negative apex of the other. The hydrogen atom is not lost by the molecule at the left (which is called the hydrogen donor) or gained by the molecule at the right (the hydrogen acceptor). The intermediate hydrogen is about a third of the way between one oxygen and the next and is said to be shared by the two molecules, even though it is still more strongly held by the donor.

Although this picture of a hydrogen bond serves well, I should point out that no mathematical model yet devised has proved to be completely successful in explaining the bonding. The interaction of two water molecules is truly not the simple static one I have described. In bulk water the stablest arrangement of the molecules arises when each molecule has hydrogen bonds with four adjacent molecules. To two of them the central molecule acts as a hydrogen donor. To the other two it acts as an acceptor. Although the arrangement is the stablest one, the hydrogen bonds can still be stretched, rotated and broken. Without this flexibility water could not flow. Indeed, the arrangement of molecules is not static even when water lies undisturbed in a container. Bonds are constantly being broken and re-formed.

Water frequently dissolves a compound because the water molecule is small enough to move close to the compound's molecules. If the compound has sites with a net charge, the water forms ionic bonds through an electrostatic attraction as it presents either its oxygen atom or a hydrogen atom to the charged site. At places on the molecule where there are polar groups the water forms hydrogen bonds with the compound.

In general only oxygen and nitrogen atoms participate in hydrogen bonding. A carbon atom cannot do so because it tends to surround itself with hydrogen (to form a hydrocarbon group) and does not leave an exposed end to which a water molecule could bond. Moreover, it does not set up an electric moment to form a polar group. This difference in bonding capabilities between carbon



The effect of ammonium hydroxide on paraffin and oleic acid

and oxygen is the main reason the hydrocarbon groups in oleic acid are hydrophobic whereas the COOH group at the end of the molecule is hydrophilic.

Additional bonding can be provided (although rarely) by the natural dissociation of water into H+ and OH− ions. They are attracted to sites on the compound that have an electric moment or a net charge. Finally, the compound might be able to dissociate to provide a hydrogen atom for a water molecule. Then the negative site left on the compound attracts the positive water ion.

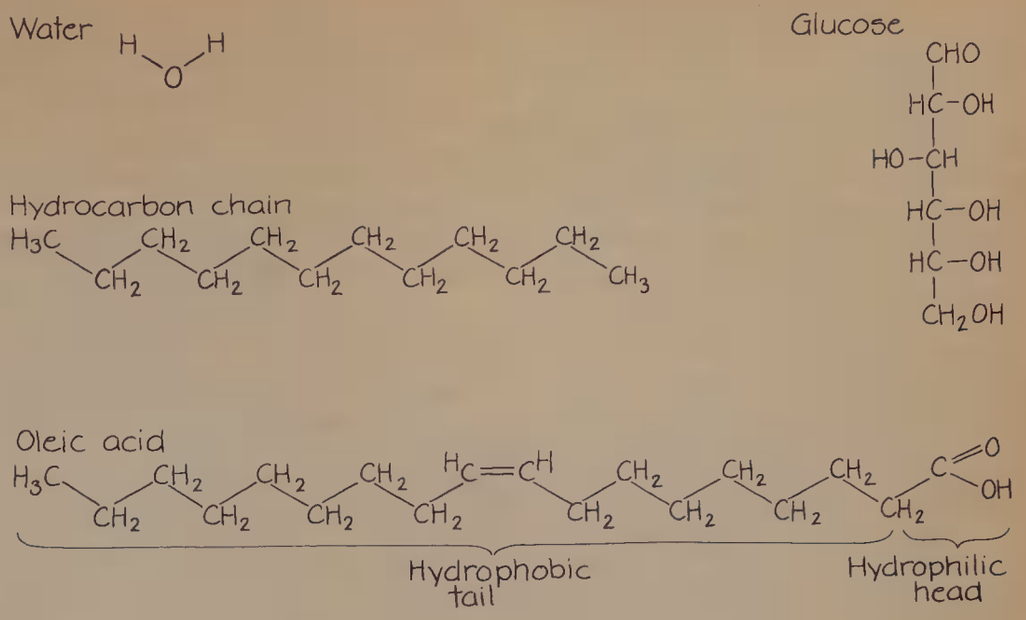
When two spherical and neutral molecules are close, there should be no electrical interaction. Because they are spheres, the center of each molecule's negative charge coincides with the center of its positive charge and so they lack electric moments. Ionic attraction is absent. Still, they can attract each other through a curious circumstance.

In the simple quantum-mechanical picture of the molecules the orbits of the electrons become synchronized, creating instantaneous electric moments that set up fleeting electric fields. The fields give rise to an attraction between the molecules. The attraction is usually termed the van der Waals force.

In a simple picture of the van der Waals force one first imagines the electrons in orbit in adjacent molecules. Each molecule should "see" the other as electrically neutral because the positive and negative charges within the molecules are equal. When the molecules are close enough, however, the two groups of electrons interact and develop synchronized orbiting.

One can then imagine that at a given instant a molecule has a momentary separation of charge, giving it an electric moment. The same is true of the other molecule. For that instant the two molecules attract each other. The attraction may fluctuate as time passes, but over time the average of the attraction is not zero; indeed, it is strong enough to hold the molecules together. Paraffin oil is held together by van der Waals forces.

Armed with this handful of models for intermolecular forces, Wissing was able to explain his experiments and to propose new ones. When he put a drop of paraffin oil on water, there was essentially no bonding between the hydrocarbons in the molecule and the water molecules. Hydrogen bonding is impossible between the two types of molecules. Other types of bonding are too uncommon to compete against the hydrogen bonding maintaining the cohesion of the water or the van der Waals force maintaining the cohesion of the oil drop. Hence the oil drop is merely pulled downward into the water surface slightly by gravity. When the drop is placed on the other solutions, the same thing happens because there is still little chance of bonding at the interface.



Some key molecular structures

The story behind the behavior of the drops of oleic acid now becomes clearer. When a drop of it is deposited on pure water, bonding immediately begins at the interface. The hydrophilic end of the oleic acid molecule is attracted to the water, but the opposite end and much of the length of the molecule are not. The hydrophilic end is commonly referred to as the head of the molecule and the opposite end is called the tail. The molecules of oleic acid at the interface rotate so that the heads line up toward the underlying water surface, leaving the tails upward.

Two types of bonding develop between the head and the water. Since the head and a water molecule are both polar, their electric moments end up in a configuration that leads to attraction. (More properly the attraction should be described as hydrogen bonding. A water molecule presents one of its hydrogen atoms to the exposed oxygen atom in the head in order to form the bond.)

The other kind of bonding involves the dissociation of the head group, which loses a hydrogen ion (a proton) to a water molecule. The loss leaves the head negatively charged (COO−) and the newly formed water ion positively charged (H_3O+). The two attract each other with ionic bonds because of their opposite electric charges.

Further bonding then begins along the interface. When the oleic acid molecules become oriented with their heads toward the water, they bond to each other in an orderly way. Adjacent heads form a hydrogen bond: the OH in one head aligns with the oxygen in a neighboring head. More bonds form between the hydrocarbons in the chain that makes up the rest of the molecule. At some sites the adjacent lengths of hydrocarbon attract each other through the van der Waals force.

The oleic acid at the edge of the drop bonds to the water as the acid molecules rotate into formation. Through the van der Waals force their motion pulls fresh oleic acid molecules onto the water, where they bond and then pull still more molecules from the bulk of the drop. The drop thins as more of its liquid moves to the edge and into molecular alignment. Soon it becomes an invisible sheet one molecule thick. The sheet cannot be stretched beyond this stage because of the bonding between its adjacent, aligned oleic acid molecules.

When a small amount of oleic acid is deposited on water, only a portion of the water surface is covered by the monolayer. With excess oleic acid the monolayer forms and the excess is left in a lenslike drop (or in several drops) on the surface. When the surface of the water has become covered with oleic acid molecules, no more of them form bonds with the water. They are left with only their van der Waals forces and the other electrical interactions that make them cohere. These surplus molecules bond themselves into the leftover drops.

When a drop of oleic acid is put on dilute hydrochloric acid, the process is much the same except for one thing. This time the abundance of hydrogen and chlorine ions in the bulk liquid precludes any systematic dissociation at the head of the oleic acid molecules. The molecules still rotate into alignment with heads down and tails up, but now their bonds with the water are almost all hydrogen bonds. Again the drop of oleic acid spreads over the surface.

It is a different story when a drop of oleic acid is placed on a dilute solution of sodium hydroxide. The solution neutralizes the acid, yielding sodium oleate. At the concentration in Wissing's experiment the oleate molecules form into tiny drops called micelles. The molecules at the interface between a micelle and the surrounding water have their

polar heads outward and their hydrocarbon tails inward. The heads bond with the surrounding water through hydrogen bonds and sometimes also through an electrostatic attraction created when a head group loses a hydrogen ion to a water molecule.

The formation of micelles can be seen clearly in Wissing's experiment. Soon after a drop of oleic acid falls onto the surface of the sodium hydroxide the drop sends out surface streams that gradually disappear. At the same time micelles form just below the streams and the drop. The bulk solution gets turbid as the micelles obstruct the passage of light through the solution.

Ammonium hydroxide also neutralizes oleic acid, but no micelles form. As the oleic acid is neutralized it is vigorously expelled from the drop in streams. The drop and the streams move rapidly over the surface. I do not understand why the acid is expelled.

The next set of experiments done by Wissing fascinates me. He began with clean water to which three or four drops of oleic acid were added. The monolayer again formed and the excess oleic acid was left in a large drop. In the photograph at the top left on page 165 the drop rests on the water, which is still shimmering after the sudden formation of the monolayer.

Away from the drop of oleic acid Wissing deposited five drops of paraffin oil. When the first one was added, the drop of oleic acid developed a broad extension. After the fifth drop of paraffin oil was put into the dish the oleic acid drop sent a stream over to the nearest paraffin drop and began to consume it. For several minutes the oleic acid slid over to the drop of oil, climbing it and eventually dissolving into it, turning the oil pink. The oleic acid drop was then fatter and more spread out.

The photograph at the middle left on page 165 shows a second drop of paraffin oil just before it was consumed by the oleic acid drop. Traces of the first drop (already consumed) can still be seen on one side of the oleic acid drop. The photograph at the bottom left was made after a second and third drop had been eaten by the oleic acid drop. The rest of the experiment continues in the photographs running from top to bottom at the right.

Neither Wissing nor I can fully explain this demonstration. He suggests that the presence of the paraffin oil offers a means whereby the hydrocarbon groups of the excess molecules of oleic acid can bond with the hydrocarbon groups in the oil. My guess is that initially the molecules in the oil drop tug on the tails of the oleic acid molecules in the surrounding monolayer. Since the molecules of the monolayer are held together, the tug is transmitted to the oleic acid drop, with the result that part of the drop is pulled toward the oil drop.

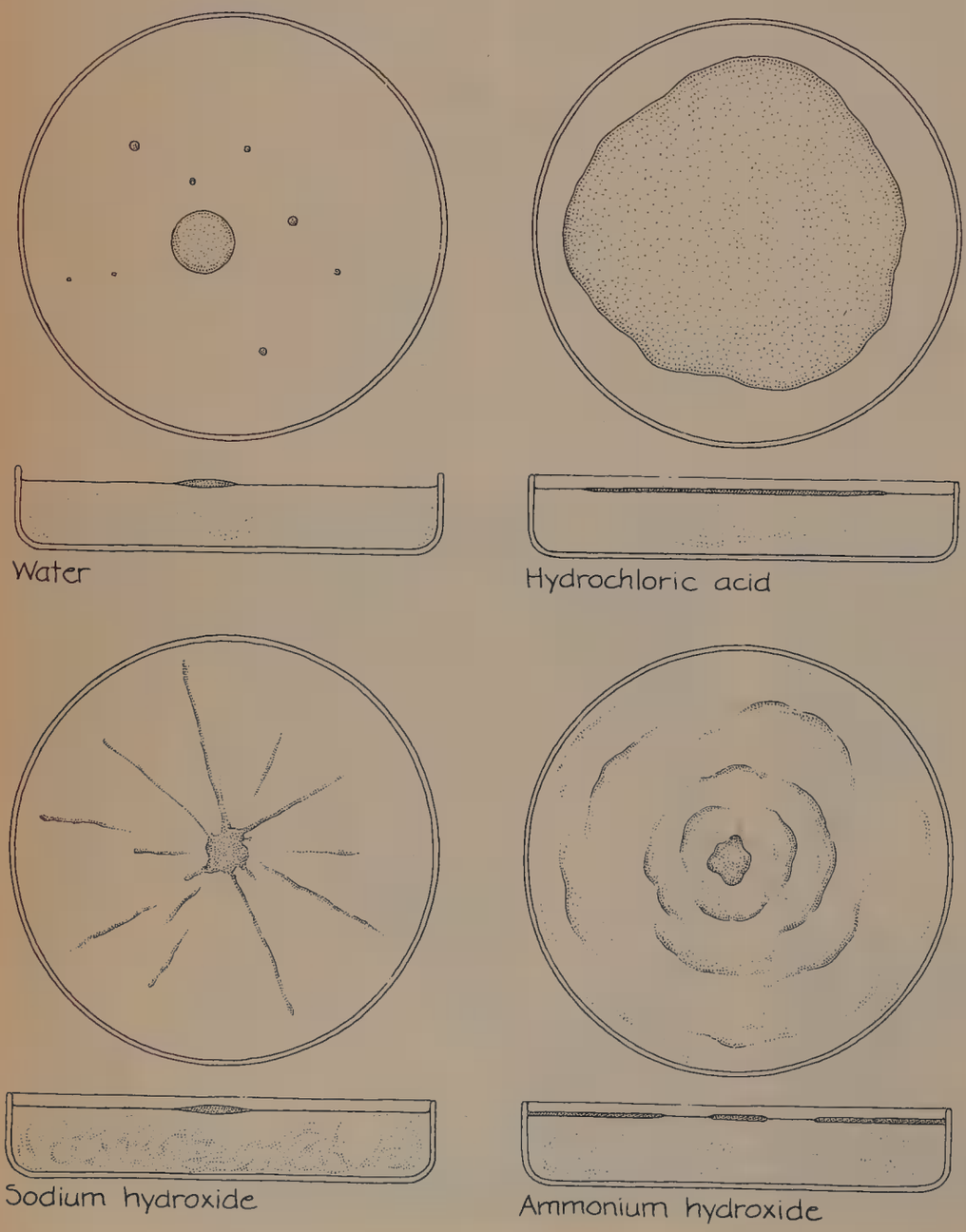
I am greatly puzzled by this experiment. Why does one drop of oil cause the drop of oleic acid to extend whereas five oil drops make the oleic acid start consuming the paraffin? I think the reason has to do with the horizontal force on the monolayer from the drops of paraffin. Suppose that before the first drop of paraffin is put in place the dish is covered with only a monolayer and the drop of excess oleic acid. Because the liquid in the drop of acid tends to slide off to the sides, it exerts a horizontal force on the monolayer.

When the first oil drop is added, it creates more force on the monolayer. Four additional drops of oil create even more force. By now an oil drop has begun to bond with the nearby oleic acid molecules surrounding it.

From which direction will a fresh batch of oleic acid come? Preferentially from the direction of the oleic acid drop because of the force with which it pushes on the monolayer. The drop of oleic acid droops onto the monolayer. This extension is then pulled by the monolayer (because of the bonding at the oil drop) and pushed by the hydrostatic pressure of the fluid in the rest of the oleic acid drop. Eventually the extension reaches and then devours the drop of oil.

Wissing sometimes changes the experiment while the oil drops are being gobbled up by the oleic acid. He pours one or two milliliters of concentrated ammonium hydroxide near one side of the oleic acid drop. The drop begins to dance wildly as the new compound neutralizes it. Micelles begin to form. The scene on the classroom screen is vivid with red drops and streams in crazy motion until the abundance of micelles blocks the passage of light through the petri dish.

Wissing's final experiments are just as entertaining. In one of them he deposits enough oleic acid on water so that a



The action of oleic acid drops on four solutions

drop of the acid remains. He puts on the middle of it a drop of Triton X-100, a commercial detergent. When the detergent has had a chance to penetrate the acid drop and reach the water, the surface of the solution is suddenly cleared of the oleic acid.

In another experiment Wissing deposits a tiny drop of oleic acid on water and adds enough paraffin oil to make an oil drop about three centimeters in diameter. Then additional oleic acid placed near the paraffin oil encircles it, forming a complicated pattern of drops that continue to move for a long time. In a separate demonstration Wissing shows how a drop of oleic acid on water can be maneuvered when a cotton swab wet with concentrated ammonium hydroxide is brought nearby. With a sudden jab of the swab he makes the drop of oleic acid jerk across the water surface. When he corrals it near the side of the dish, the drop oscillates in the vapor given off by the swab.

The formation of a monolayer by partially soluble organic compounds played a major role in the early work on determining the size of molecules. In 1890 Lord Rayleigh employed a minute amount of olive oil to estimate the size of its molecules. A fine platinum wire was dipped in the oil and then weighed on a sensitive balance. Some of the oil was released onto a circular water surface 84 centimeters in diameter and the wire was reweighed. The difference in the two measurements was the weight of the oil then covering the water surface.

To monitor the extent of spreading by the monolayer Rayleigh sprinkled the water with fine grains of camphor. As camphor dissolves in water on one side of a grain, it decreases the surface tension of the water. The larger surface tension on the other side jerks the grain. When Rayleigh added olive oil, the surface tension was too low for the camphor dance. He could therefore monitor the extent of the oil layer by watching the camphor grains. Adding just enough oil to stop all the grains from dancing, he then had the water surface covered evenly with a monolayer. From the surface area and the weight of the drop he computed the thickness of the layer as being about 17 angstrom units. Rayleigh thought this number was also the length of the molecules forming the monolayer.

Richard E. Crandall and Jean F. Delord of Reed College have developed a modified form of Rayleigh's experiment for their students in introductory physics. In the exercise a student calculates the length of the oleic acid molecule, the average bond length along the molecule and the mass of the carbon atom. The first step is to sprinkle a water surface with lycopodium powder. When a drop of oleic acid is added, the monolayer pushes the powder outward into a circle



The Best of Christmas from Questar

From the first moment we started making Questar telescopes, early in the fifties, we took a fierce pride in our quality. We believed that the magic inherent in a carefully executed Maksutov design would enable us to make superior instruments that were also conveniently portable, and we were determined to settle for nothing less. Now, more than thirty years later, Questar still has the reputation, carefully nurtured over the years, for making the world's finest. And our skill in designing instruments for special purposes has grown, culminating in the I-R 100 award to Questar Corporation by Industrial Research and Development magazine for one of the 100 most significant technical products of 1983—the QM 1.

But to go back to the instrument that started it all, the Questar® 3½: this is the telescope with the amazing versatility that can take you to the far reaches of the universe or let you explore from your comfortable armchair the astonishing detail of rock, tree or bird. And its quality continues, with the finest optics that can be produced combined with careful attention to every mechanical part, in a size so compact that you can carry it in a fine luggage case wherever you go.

For a family with a variety of interests, a Questar is the perfect companion. Shared by a dedicated astronomer, a general observer who delights in all natural phenomena, or a photographer of things seldom, if ever, seen with the naked eye, a Questar will never be idle. Why don't you treat yourselves to this truly unique Christmas gift?

• • •

Hubert Entrop uses both his Questar 3½ and 7 for photographing the deep-sky objects that we feature here from time to time. The magnificent photograph of Omega Centauri above, done with the 7, is reproduced from a 5x7 print enlargement of a 35 mm Ektachrome 200 slide. He writes that the air was extremely rough, so that the guide star was nothing more than a fuzzy ball. In spite of this the round star images indicate that the Questar was guiding right on course. Exposure time one hour at f13.5.

• • •

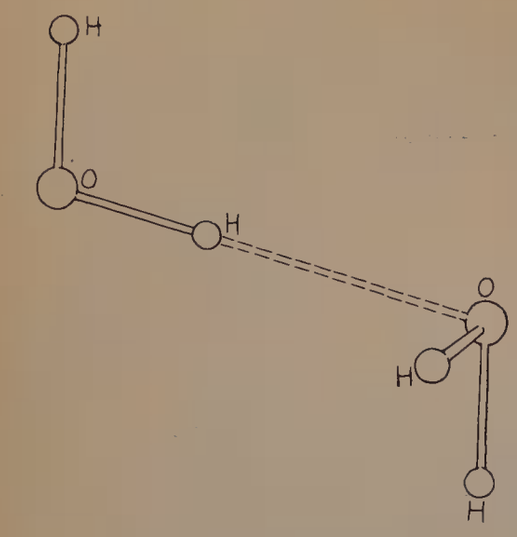
A collection of great photographs of the moon taken with the Questar 3½ and 7 is brought together in a book, THE QUESTAR MOON, which will make a beautiful addition to your astronomy library. Format size is 11x14 inches. Pages are perforated to be easily removable for framing, if desired, without destroying the soft-cover book. The resolution of detail in these photographs is so remarkable that they invite comparison with similar views taken with the 24" refractor at Pic du Midi observatory, published in Kopal's "Photographic Atlas Of The Moon." Price $9.95.

©Questar Corporation 1983.

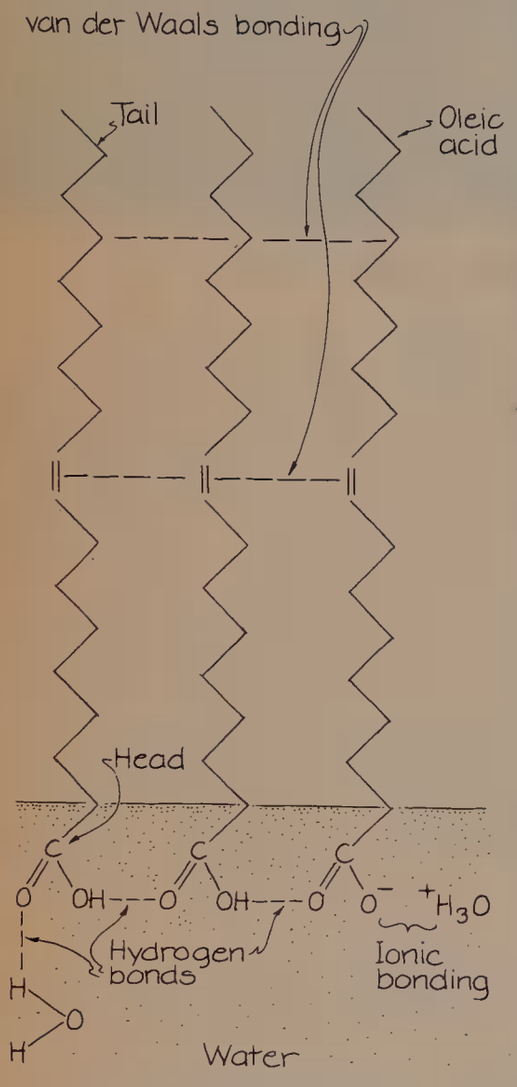
LET US SEND YOU OUR LITERATURE DESCRIBING QUESTAR, THE WORLD'S FINEST MOST VERSATILE TELESCOPE. PLEASE SEND $2 TO COVER MAILING COSTS ON THIS CONTINENT. BY AIR TO SOUTH AMERICA, $3.50; EUROPE AND NORTH AFRICA, $4; ELSEWHERE $4.50. INQUIRE ABOUT OUR EXTENDED PAYMENT PLAN.

QUESTAR

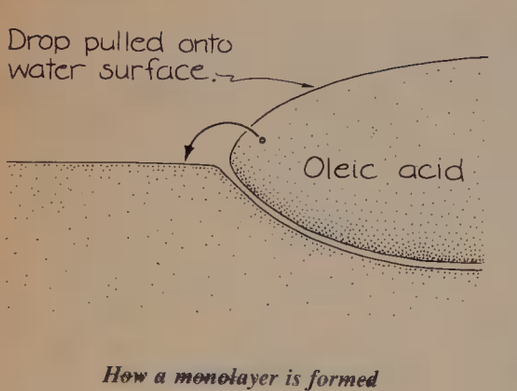
P.O. BOX C, Dept. 207, NEW HOPE, PA 18938
PHONE (215) 862-5277



The hydrogen bond of water



A monolayer of oleic acid



How a monolayer is formed

and then contracts into an irregular perimeter. Before the circle contracts the student sketches its boundary on graph paper. Later the sketch is measured to ascertain the area of the circle.

The student is given several facts. The solution of oleic acid is prepared with 200 parts per million of acid in alcohol, measured in terms of volume. The size of a drop of acid is approximately .05 cubic centimeter. When the molecules of acid are aligned with heads down and tails up, a head occupies an area that can be approximated as a square with a side .1 the length of the molecule. The masses of O, C and H are in the ratio of 16 : 12 : 1. The density of oleic acid is about .895 gram per cubic centimeter.

By multiplying the concentration of the oleic acid and the volume of the drop the student calculates the volume of oleic acid that spreads over the water. Dividing the volume of the acid by the surface area of the monolayer yields the height of the layer. That height is also the length of the oleic acid molecule.

The next step is to estimate the bond length, that is, the distance between two adjacent carbon atoms along the length of the molecule. For simplicity you can consider the double bond at the center of the molecule as a single bond; the rest of the bonds along the length of the hydrocarbon chain are true single bonds. You also might ignore the nonlinear structure at the head of the molecule. In this way you are dealing with 17 bonds between carbon atoms along the molecule, or 18 if the hydrogen at the tail is included. Assume that the bonds lie along a straight line. When their number is divided into the length of the molecule, you find that the bond length is on the order of one angstrom.

Multiplying the volume of oleic acid by the density of the acid gives the mass of the monolayer. Find the mass of each molecule by means of the estimate that the head of the molecule is a square with a side .1 the length of the molecule. Since the molecule's length is known by now, the area of the square of a head is easy to compute.

The number of oleic acid molecules is computed by dividing the area of the monolayer by the area of each head. Dividing the mass of the layer by the number of molecules yields the mass of each molecule. Finally, the mass of a single carbon atom is determined by counting the number of each type of atom in the molecule and employing the known ratios of their masses.

Molecules are notoriously difficult to picture because of their smallness and complexity. Yet from Rayleigh's experiments one can estimate their size. From Crandall and Delord's laboratory exercise one can determine bond length. With Wissing's colorful demonstrations one can even imagine the ordering of the molecules along the interface between water and an organic compound and can make sense of why some organic compounds spread over water and others do not.

STATEMENT OF OWNERSHIP, MANAGEMENT AND CIRCULATION (required by 39 U.S.C. 3685). *1A.* Title of publication: Scientific American. *1B.* Publication number: 00368733. *2.* Date of filing: September 27, 1983. *3.* Frequency of issue: monthly. *3A.* Number of issues published annually: 12. *3B.* Annual subscription price: U.S. and its possessions, 1 year $24; all other countries, 1 year $33. *4.* Complete mailing address of known office of publication: 415 Madison Avenue, New York, N.Y. 10017. *5.* Complete mailing address of the headquarters or general business offices of the publisher: 415 Madison Avenue, New York, N.Y. 10017. *6.* Full names and complete mailing address of publisher, editor, and managing editor: Publisher, Gerard Piel, 415 Madison Avenue, New York, N.Y. 10017. Editor, Dennis Flanagan, 415 Madison Avenue, New York, N.Y. 10017. Managing Editor, none. *7.* Owner: Scientific American, Inc., 415 Madison Avenue, New York, N.Y. 10017; Atwell & Co. (for Estate of John Hay Whitney), P.O. Box 456, Wall Street Station, New York, N.Y. 10005; Buckley & Co. (for Elinor G. Mannucci), c/o Fiduciary Trust Co. of New York, P.O. Box 3199, Church Street Station, New York, N.Y. 10008; George S. Conn, 20 Bittersweet Lane, Mount Kisco, N.Y. 10549; Cudd & Co. for (a) Helen Rosenwald Snellenburg Trust, (b) Robert L. Rosenwald Sr. Trust, (c) Trust U/W Lessing J. Rosenwald for benefit of Edith G. Rosenwald and children, (d) I. M. Scott, The Chase Manhattan Bank, 1 Chase Manhattan Plaza, New York, N.Y. 10015; Martin M. Davidson, 11 River Lane, Westport, Conn. 06880; Dengel & Co. (for Nathan W. Levin), c/o Fiduciary Trust Co. of New York, P.O. Box 3199, Church Street Station, New York, N.Y. 10015; Richard E. Deutsch, c/o The Chase Manhattan Bank, P.O. Box 1508, Church Street Station, New York, N.Y. 10008; Egger & Co. (for Bayard Ewing), c/o The Chase Manhattan Bank, P.O. Box 1508, Church Street Station, New York, N.Y. 10008; Dorothy D. Eweson, c/o Brunt & Co., P.O. Box 6626, Journal Square, Jersey City, N.J. 07306; Excelsior Fund, 545 Madison Avenue, New York, N.Y. 10022; Dennis Flanagan, c/o Scientific American, Inc., 415 Madison Avenue, New York, N.Y. 10017; William T. Golden, Room 4201, 40 Wall Street, New York, N.Y. 10005; C. John Kirby, c/o Scientific American, Inc., 415 Madison Avenue, New York, N.Y. 10017; Frasier W. McCann, Suite 1706, 680 Fifth Avenue, New York, N.Y. 10019; Donald H. Miller, Jr., c/o Scientific American, Inc., 415 Madison Avenue, New York, N.Y. 10017; Eleanor Jackson Piel, Jr., 320 Central Park West, New York, N.Y. 10025; Gerard Piel, c/o Scientific American, Inc., 415 Madison Avenue, New York, N.Y. 10017; Jonathan Piel, c/o Scientific American, Inc., 415 Madison Avenue, New York, N.Y. 10017; The President & Fellows of Harvard College Trust, 70 Federal Street, Boston, Mass. 02110; Barbara K. Rosenwald, P.O. Box 3625, Norfolk, Va. 23514; Julius Rosenwald II, c/o The Chase Manhattan Bank, P.O. Box 1508, Church Street Station, New York, N.Y. 10008; Rowe & Co. (for Rhode Island Charities Trust), c/o Industrial National Bank of Rhode Island, Trust Department, 100 Westminster Street, Providence, R.I. 02903; Stanley Schaefer, 418 Yerba Santa Avenue, Los Altos, Calif. 94022; Joan R. Scott, c/o The Chase Manhattan Bank, P.O. Box 1508, Church Street Station, New York, N.Y. 10008. *8.* Known bondholders, mortgagees, and other security holders owning or holding 1 percent or more of total amount of bonds, mortgages or other securities: none. *9.* For completion by nonprofit organizations authorized to mail at special rates (Section 423.12, DMM only). The purpose, function, and nonprofit status of this organization and the exempt status for Federal income tax purposes: Not applicable to this organization. *10.* Extent and nature of circulation: *A.* Total number of copies (net press run): average number of copies each issue during preceding 12 months, 797,619; actual number of copies of single issue published nearest to filing date, 841,615. *B.* Paid circulation: *1.* Sales through dealers and carriers, street vendors and counter sales: average number of copies each issue during preceding 12 months, 149,573; actual number of copies of single issue published nearest to filing date, 201,626. *2.* Mail subscription: average number of copies each issue during preceding 12 months, 524,073; actual number of copies of single issue published nearest to filing date, 503,800. *C.* Total paid circulation (sum of *10B1* and *10B2*): average number of copies each issue during preceding 12 months, 673,646; actual number of copies of single issue published nearest to filing date, 705,426. *D.* Free distribution by mail, carrier or other means, samples, complimentary, and other free copies: average number of copies each issue during preceding 12 months, 6,784; actual number of copies of single issue published nearest to filing date, 7,128. *E.* Total distribution (sum of *C* and *D*): average number of copies each issue during preceding 12 months, 680,430; actual number of copies of single issue published nearest to filing date, 712,554. *F.* Copies not distributed: *1.* Office use, leftover, unaccounted, spoiled after printing: average number of copies each issue during preceding 12 months, 6,614; actual number of copies of single issue published nearest to filing date, 6,394. *2.* Return from news agents: average number of copies each issue during preceding 12 months, 110,575; actual number of copies of single issue published nearest to filing date, 122,667. *G.* Total (sum of *E, F1* and *2*—should equal net press run shown in *A*): average number of copies each issue during preceding 12 months, 797,619; actual number of copies of single issue published nearest to filing date, 841,615. *11.* I certify that the statements made by me above are correct and complete. (Signed) George S. Conn, Vice-President and Gen. Manager.



Test Drive the Leica R4.

We invite you to experience the perfection of the Leica® R4. We invite you to hold, inspect, and then test the Leica that has been acclaimed "...a miracle of modern technology." A limited production SLR of astonishing versatility, precision and beauty. With flawless Leitz® optics, two exposure measurement systems, five operation modes, and a series of advances unequalled in the history of photography.

The Leica R4 is Leica's proudest achievement. A photographic instrument designed for perfection, not price. Like every Leica since the beginning, it is lavished with attention to detail. Hand finishing. Hand assembly. Hand testing. No shortcuts or compromises. Because the Leica R4 is designed and built never to fail. And to have no limit to its useful life.

We invite you to experience the Leica R4. And through it enter Leica's world of superior photography. To see and appreciate for yourself the perfection of the photographic image that is uniquely Leica's.

You can test drive the Leica R4 now, at your local Leica dealer. And to show his appreciation for your visit and your interest, your Leica dealer will enter your name in a special drawing for the ultimate sports coupe—the BMW 633CSi.* We think you'll find this high performance German import, the perfect accessory to your new Leica.

For the name of your nearest Leica dealer write E. Leitz, Inc., 24 Link Drive, Rockleigh, N.J. 07647 or call (800) 223-0507, ext. 801. In New Jersey call (800) 223-0509, ext. 801.

And you could win a BMW 633 CSi to drive it home in.

*No purchase necessary. Entries must be made through your Leica dealer before January 31, 1984. The winner will be notified by March 30, 1984. Void where prohibited by law

ANNUAL INDEX

The following index lists all the authors and articles that appeared in SCIENTIFIC AMERICAN during 1983. Also indexed are "Metamagical Themas," "Mathematical Games," "Computer Recreations" and "The Amateur Scientist."

AUTHORS

Abraham, Eitan, Colin T. Seaton and S. Desmond Smith. THE OPTICAL COMPUTER; February, page 85.

Alkon, Daniel L. LEARNING IN A MARINE SNAIL; July, page 70.

Anders, Edward, and Roy S. Lewis. INTERSTELLAR MATTER IN METEORITES; August, page 66.

Angell, James B., Stephen C. Terry and Phillip W. Barth. SILICON MICROMECHANICAL DEVICES; April, page 44.

Aune, Petter, Ronald L. Sack and Arne Selberg. THE STAVE CHURCHES OF NORWAY; August, page 96.

Barth, Phillip W., James B. Angell and Stephen C. Terry. SILICON MICROMECHANICAL DEVICES; April, page 44.

Batie, Sandra S., and Robert G. Healy. THE FUTURE OF AMERICAN AGRICULTURE; February, page 45.

Bergerud, Arthur T. PREY SWITCHING IN A SIMPLE ECOSYSTEM; December, page 130.

Bertsch, George F. VIBRATIONS OF THE ATOMIC NUCLEUS; May, page 62.

Bickerton, Derek. CREOLE LANGUAGES; July, page 116.

Bigelow, Charles, and Donald Day. DIGITAL TYPOGRAPHY; August, page 106.

Birchall, J. D., and Anthony Kelly. NEW INORGANIC MATERIALS; May, page 104.

Blair, Douglas H., and Robert A. Pollak. RATIONAL COLLECTIVE CHOICE; August, page 88.

Blechman, Barry M., and Mark R. Moore. A NUCLEAR-WEAPON-FREE ZONE IN EUROPE; April, page 37.

Blodgett, Jr., Albert J. MICROELECTRONIC PACKAGING; July, page 86.

Bogucki, Peter, and Ryszard Grygiel. EARLY FARMERS OF THE NORTH EUROPEAN PLAIN; April, page 104.

Bonner, John Tyler. CHEMICAL SIGNALS OF SOCIAL AMOEBAE; April, page 114.

Borgese, Elisabeth Mann. THE LAW OF THE SEA; March, page 42.

Broecker, Wallace S. THE OCEAN; September, page 146.

Bunn, Matthew, and Kosta Tsipis. THE UNCERTAINTIES OF A PREEMPTIVE NUCLEAR ATTACK; November, page 38.

Bur, Michel. THE SOCIAL INFLUENCE OF THE MOTTE-AND-BAILEY CASTLE; May, page 132.

Burchfiel, B. Clark. THE CONTINENTAL CRUST; September, page 130.

Burns, Jack O., and R. Marcus Price. CENTAURUS A: THE NEAREST ACTIVE GALAXY; November, page 56.

Chilton, Mary-Dell. A VECTOR FOR INTRODUCING NEW GENES INTO PLANTS; June, page 50.

Cloud, Preston. THE BIOSPHERE; September, page 176.

Conn, Robert W. THE ENGINEERING OF MAGNETIC FUSION REACTORS; October, page 60.

Courtillot, Vincent, and Gregory E. Vink. HOW CONTINENTS BREAK UP; July, page 42.

Daly, John W., and Charles W. Myers. DART-POISON FROGS; February, page 120.

Darnell, Jr., James E. THE PROCESSING OF RNA; October, page 90.

Dauben, Joseph W. GEORG CANTOR AND THE ORIGINS OF TRANSFINITE SET THEORY; June, page 122.

Day, Donald, and Charles Bigelow. DIGITAL TYPOGRAPHY; August, page 106.

DeAre, Diana, and Larry Long. THE SLOWING OF URBANIZATION IN THE U.S.; July, page 33.

de Duve, Christian. MICROBODIES IN THE LIVING CELL; May, page 74.

DeJong, Gerben, and Raymond Lifchez. PHYSICAL DISABILITY AND PUBLIC POLICY; June, page 40.

De Kepper, Patrick, Irving R. Epstein, Kenneth Kustin and Miklós Orbán. OSCILLATING CHEMICAL REACTIONS; March, page 112.

DeWitt, Bryce S. QUANTUM GRAVITY; December, page 112.

Diaconis, Persi, and Bradley Efron. COMPUTER-INTENSIVE METHODS IN STATISTICS; May, page 116.

Dickerson, Richard E. THE DNA HELIX AND HOW IT IS READ; December, page 94.

Dicus, Duane A., John R. Letaw, Doris C. Teplitz and Vigdor L. Teplitz. THE FUTURE OF THE UNIVERSE; March, page 90.

Döhler, Gottfried H. SOLID-STATE SUPERLATTICES; November, page 144.

Donaldson, Lauren R., and Timothy Joyner. THE SALMONID FISHES AS A NATURAL LIVESTOCK; July, page 50.

Edmond, John M., and Karen Von Damm. HOT SPRINGS ON THE OCEAN FLOOR; April, page 78.

Efron, Bradley, and Persi Diaconis. COMPUTER-INTENSIVE METHODS IN STATISTICS; May, page 116.

Epstein, Irving R., Kenneth Kustin, Patrick De Kepper and Miklós Orbán. OSCILLATING CHEMICAL REACTIONS; March, page 112.

Fairservis, Jr., Walter A. THE SCRIPT OF THE INDUS VALLEY CIVILIZATION; March, page 58.

Feder, Joseph, and William R. Tolbert. THE LARGE-SCALE CULTIVATION OF MAMMALIAN CELLS; January, page 36.

Fletcher, Neville H., and Suszanne Thwaites. THE PHYSICS OF ORGAN PIPES; January, page 94.

Francheteau, Jean. THE OCEANIC CRUST; September, page 114.

Francis, Peter. GIANT VOLCANIC CALDERAS; June, page 60.

Francis, Peter, and Stephen Self. THE ERUPTION OF KRAKATAU; November, page 172.

Gat, Joel R., and Ilana Steinhorn. THE DEAD SEA; October, page 102.

Geller, Stephen A. AUTOPSY; March, page 124.

Grivell, Leslie A. MITOCHONDRIAL DNA; March, page 78.

Gross, Albert C., Chester R. Kyle and Douglas J. Malewicki. THE AERODYNAMICS OF HUMAN-POWERED LAND VEHICLES; December, page 142.

Grygiel, Ryszard, and Peter Bogucki. EARLY FARMERS OF THE NORTH EUROPEAN PLAIN; April, page 104.

Harari, Haim. THE STRUCTURE OF QUARKS AND LEPTONS; April, page 56.

Harbron, John D. MODERN ICEBREAKERS; December, page 49.

Healy, Robert G., and Sandra S. Batie. THE FUTURE OF AMERICAN AGRICULTURE; February, page 45.

Hoffman, Donald D. THE INTERPRETATION OF VISUAL ILLUSIONS; December, page 154.

Hua, Jue-ming. THE MASS PRODUCTION OF IRON CASTINGS IN ANCIENT CHINA; January, page 120.

Hudspeth, A. J. THE HAIR CELLS OF THE INNER EAR; January, page 54.

Ingersoll, Andrew P. THE ATMOSPHERE; September, page 162.

Jeanloz, Raymond. THE EARTH'S CORE; September, page 56.

Johnson, Torrence V., and Laurence A. Soderblom. IO; December, page 56.

Joyner, Timothy, and Lauren R. Donaldson. THE SALMONID FISHES AS A NATURAL LIVESTOCK; July, page 50.

Kanwisher, John W., and Sam H. Ridgway. THE PHYSIOLOGICAL ECOLOGY OF WHALES AND PORPOISES; June, page 110.

Kaplan, Donald R. THE DEVELOPMENT OF PALM LEAVES; July, page 98.

Kelly, Anthony, and J. D. Birchall. NEW INORGANIC MATERIALS; May, page 104.

Kustin, Kenneth, Irving R. Epstein, Patrick De Kepper and Miklós Orbán. OSCILLATING CHEMICAL REACTIONS; March, page 112.

Kyle, Chester R., Albert C. Gross and Douglas J. Malewicki. THE AERODYNAMICS OF HUMAN-POWERED LAND VEHICLES; December, page 142.

Leedy, Jr., Walter C. FAN VAULTING; February, page 134.

Lerner, Richard A. SYNTHETIC VACCINES; February, page 66.

Letaw, John R., Duane A. Dicus, Doris C. Teplitz and Vigdor L. Teplitz. THE FUTURE OF THE UNIVERSE; March, page 90.

Lewis, Roy S., and Edward Anders. INTERSTELLAR MATTER IN METEORITES; August, page 66.

Lifchez, Raymond, and Gerben DeJong. PHYSICAL DISABILITY AND PUBLIC POLICY; June, page 40.

Lin, Shenglong, and Wu Rukang. PE-

KING MAN; June, page 86.
Long, Larry, and Diana DeAre. THE SLOWING OF URBANIZATION IN THE U.S.; July, page 33.
Malewicki, Douglas J., Albert C. Gross and Chester R. Kyle. THE AERODYNAMICS OF HUMAN-POWERED LAND VEHICLES; December, page 142.
Margon, Bruce. THE ORIGIN OF THE COSMIC X-RAY BACKGROUND; January, page 104.
Martin, Philip L. LABOR-INTENSIVE AGRICULTURE; October, page 54.
McCloskey, Michael. INTUITIVE PHYSICS; April, page 122.
McKenzie, D. P. THE EARTH'S MANTLE; September, page 66.
Mistry, Nariman B., Ronald A. Poling and Edward H. Thorndike. PARTICLES WITH NAKED BEAUTY; July, page 106.
Moore, Mark R., and Barry M. Blechman. A NUCLEAR-WEAPON-FREE ZONE IN EUROPE; April, page 37.
Morrison, Adrian R. A WINDOW ON THE SLEEPING BRAIN; April, page 94.
Mossman, David J., and William A. S. Sarjeant. THE FOOTPRINTS OF EXTINCT ANIMALS; January, page 74.
Myers, Charles W., and John W. Daly. DART-POISON FROGS; February, page 120.
Orbán, Miklós, Irving R. Epstein, Kenneth Kustin and Patrick De Kepper. OSCILLATING CHEMICAL REACTIONS; March, page 112.
Parker, E. N. MAGNETIC FIELDS IN THE COSMOS; August, page 44.
Patterson, David A. MICROPROGRAMMING; March, page 50.
Pestka, Sidney. THE PURIFICATION AND MANUFACTURE OF HUMAN INTERFERONS; August, page 36.
Poling, Ronald A., Nariman B. Mistry and Edward H. Thorndike. PARTICLES WITH NAKED BEAUTY; July, page 106.
Pollak, Robert A., and Douglas H. Blair. RATIONAL COLLECTIVE CHOICE; August, page 88.
Pond, Wilson G. MODERN PORK PRODUCTION; May, page 96.
Prestwich, Glenn D. THE CHEMICAL DEFENSES OF TERMITES; August, page 78.
Price, R. Marcus, and Jack O. Burns. CENTAURUS A: THE NEAREST ACTIVE GALAXY; November, page 56.
Raibert, Marc H., and Ivan E. Sutherland. MACHINES THAT WALK; January, page 44.
Rebbi, Claudio. THE LATTICE THEORY OF QUARK CONFINEMENT; February, page 54.
Reeves, B. O. K. SIX MILLENNIUMS OF BUFFALO KILLS; October, page 120.
Renfrew, Colin. THE SOCIAL ARCHAEOLOGY OF MEGALITHIC MONUMENTS; November, page 152.
Ridgway, Sam H., and John W. Kanwisher. THE PHYSIOLOGICAL ECOLOGY OF WHALES AND PORPOISES; June, page 110.
Rosenthal, Gerald A. A SEED-EATING BEETLE'S ADAPTATIONS TO A POISONOUS SEED; November, page 164.
Rubin, Vera C. DARK MATTER IN SPIRAL GALAXIES; June, page 96.
Sack, Ronald L., Petter Aune and Arne Selberg. THE STAVE CHURCHES OF NORWAY; August, page 96.
Sarjeant, William A. S., and David J. Mossman. THE FOOTPRINTS OF EXTINCT ANIMALS; January, page 74.
Seaton, Colin T., Eitan Abraham and S. Desmond Smith. THE OPTICAL COMPUTER; February, page 85.
Selberg, Arne, Petter Aune and Ronald L. Sack. THE STAVE CHURCHES OF NORWAY; August, page 96.
Self, Stephen, and Peter Francis. THE ERUPTION OF KRAKATAU; November, page 172.
Shettleworth, Sara J. MEMORY IN FOOD-HOARDING BIRDS; March, page 102.
Shulman, R. G. NMR SPECTROSCOPY OF LIVING CELLS; January, page 86.
Siever, Raymond. THE DYNAMIC EARTH; September, page 46.
Silk, Joseph, Alexander S. Szalay and Yakov B. Zel'dovich. THE LARGE-SCALE STRUCTURE OF THE UNIVERSE; October, page 72.
Smith, S. Desmond, Eitan Abraham and Colin T. Seaton. THE OPTICAL COMPUTER; February, page 85.
Soderblom, Laurence A., and Torrence V. Johnson. IO; December, page 56.
Steinhorn, Ilana, and Joel R. Gat. THE DEAD SEA; October, page 102.
Sutherland, Ivan E., and Marc H. Raibert. MACHINES THAT WALK; January, page 44.
Szalay, Alexander S., Joseph Silk and Yakov B. Zel'dovich. THE LARGE-SCALE STRUCTURE OF THE UNIVERSE; October, page 72.
Teplitz, Doris C., Duane A. Dicus, John R. Letaw and Vigdor L. Teplitz. THE FUTURE OF THE UNIVERSE; March, page 90.
Teplitz, Vigdor L., Duane A. Dicus, John R. Letaw and Doris C. Teplitz. THE FUTURE OF THE UNIVERSE; March, page 90.
Terry, Stephen C., James B. Angell and Phillip W. Barth. SILICON MICROMECHANICAL DEVICES; April, page 44.
Thorndike, Edward H., Nariman B. Mistry and Ronald A. Poling. PARTICLES WITH NAKED BEAUTY; July, page 106.
Thwaites, Suszanne, and Neville H. Fletcher. THE PHYSICS OF ORGAN PIPES; January, page 94.
Tolbert, William R., and Joseph Feder. THE LARGE-SCALE CULTIVATION OF MAMMALIAN CELLS; January, page 36.
Trunkey, Donald D. TRAUMA; August, page 28.
Tsipis, Kosta, and Matthew Bunn. THE UNCERTAINTIES OF A PREEMPTIVE NUCLEAR ATTACK; November, page 38.
Vink, Gregory E., and Vincent Courtillot. HOW CONTINENTS BREAK UP; July, page 42.
Von Damm, Karen, and John M. Edmond. HOT SPRINGS ON THE OCEAN FLOOR; April, page 78.
Walker, Paul F. SMART WEAPONS IN NAVAL WARFARE; May, page 53.
Ward, Peter. THE EXTINCTION OF THE AMMONITES; October, page 136.
Wasp, Edward J. SLURRY PIPELINES; November, page 48.
Weinberg, Robert A. A MOLECULAR BASIS OF CANCER; November, page 126.
Weiner, Lewis. THE SLIDE FASTENER; June, page 132.
Wells, Peter S. AN EARLY IRON AGE FARM COMMUNITY IN CENTRAL EUROPE; December, page 68.
Winfree, Arthur T. SUDDEN CARDIAC DEATH: A PROBLEM IN TOPOLOGY; May, page 144.
Wolfe, Jeremy M. HIDDEN VISUAL PROCESSES; February, page 94.
Wolfson, Richard. THE ACTIVE SOLAR CORONA; February, page 104.
Wu, Rukang, and Lin Shenglong. PEKING MAN; June, page 86.
York, Herbert F. BILATERAL NEGOTIATIONS AND THE ARMS RACE; October, page 149.
Zel'dovich, Yakov B., Joseph Silk and Alexander S. Szalay. THE LARGE-SCALE STRUCTURE OF THE UNIVERSE; October, page 72.

ARTICLES

AGRICULTURE, LABOR-INTENSIVE, by Philip L. Martin; October, page 54.
AMERICAN AGRICULTURE, THE FUTURE OF, by Sandra S. Batie and Robert G. Healy; February, page 45.
AMMONITES, THE EXTINCTION OF THE, by Peter Ward; October, page 136.
ARMS RACE, BILATERAL NEGOTIATIONS AND THE, by Herbert F. York; October, page 149.
ATMOSPHERE, THE, by Andrew P. Ingersoll; September, page 162.
ATOMIC NUCLEUS, VIBRATIONS OF THE, by George F. Bertsch; May, page 62.
AUTOPSY, by Stephen A. Geller; March, page 124.
BIOSPHERE, THE, by Preston Cloud; September, page 176.
BRAIN, A WINDOW ON THE SLEEPING, by Adrian R. Morrison; April, page 94.
BUFFALO KILLS, SIX MILLENNIUMS OF, by B. O. K. Reeves; October, page 120.
CANCER, A MOLECULAR BASIS OF, by Robert A. Weinberg; November, page 126.
CANTOR AND THE ORIGINS OF TRANSFINITE SET THEORY, GEORG, by Joseph W. Dauben; June, page 122.
CENTAURUS A: THE NEAREST ACTIVE GALAXY, by Jack O. Burns and R. Marcus Price; November, page 56.
CHEMICAL REACTIONS, OSCILLATING, by Irving R. Epstein, Kenneth Kustin, Patrick De Kepper and Miklós Orbán; March, page 112.
COMPUTER-INTENSIVE METHODS IN STATISTICS, by Persi Diaconis and Bradley Efron; May, page 116.
CONTINENTAL CRUST, THE, by B. Clark Burchfiel; September, page 130.
CONTINENTS BREAK UP, HOW, by Vincent Courtillot and Gregory E. Vink; July,

page 42.
CREOLE LANGUAGES, by Derek Bickerton; July, page 116.
DART-POISON FROGS, by Charles W. Myers and John W. Daly; February, page 120.
DEAD SEA, THE, by Ilana Steinhorn and Joel R. Gat; October, page 102.
DIGITAL TYPOGRAPHY, by Charles Bigelow and Donald Day; August, page 106.
DNA HELIX AND HOW IT IS READ, THE, by Richard E. Dickerson; December, page 94.
EARLY FARMERS OF THE NORTH EUROPEAN PLAIN, by Peter Bogucki and Ryszard Grygiel; April, page 104.
EARTH, THE DYNAMIC, by Raymond Siever; September, page 46.
EARTH'S CORE, THE, by Raymond Jeanloz; September, page 56.
EARTH'S MANTLE, THE, by D. P. McKenzie; September, page 66.
FAN VAULTING, by Walter C. Leedy, Jr.; February, page 134.
FOOTPRINTS OF EXTINCT ANIMALS, THE, by David J. Mossman and William A. S. Sarjeant; January, page 74.
HAIR CELLS OF THE INNER EAR, THE, by A. J. Hudspeth; January, page 54.
HOT SPRINGS ON THE OCEAN FLOOR, by John M. Edmond and Karen Von Damm; April, page 78.
HUMAN-POWERED LAND VEHICLES, THE AERODYNAMICS OF, by Albert C. Gross, Chester R. Kyle and Douglas J. Malewicki; December, page 142.
ICEBREAKERS, MODERN, by John D. Harbron; December, page 49.
INDUS VALLEY CIVILIZATION, THE SCRIPT OF THE, by Walter A. Fairservis, Jr.; March, page 58.
INORGANIC MATERIALS, NEW, by J. D. Birchall and Anthony Kelly; May, page 104.
INTERFERONS, THE PURIFICATION AND MANUFACTURE OF HUMAN, by Sidney Pestka; August, page 36.
INTUITIVE PHYSICS, by Michael McCloskey; April, page 122.
IO, by Torrence V. Johnson and Laurence A. Soderblom; December, page 56.
IRON AGE FARM COMMUNITY IN CENTRAL EUROPE, AN EARLY, by Peter S. Wells; December, page 68.
IRON CASTINGS IN ANCIENT CHINA, THE MASS PRODUCTION OF, by Hua Jueming; January, page 120.
KRAKATAU, THE ERUPTION OF, by Peter Francis and Stephen Self; November, page 172.
LAW OF THE SEA, THE, by Elisabeth Mann Borgese; March, page 42.
LEARNING IN A MARINE SNAIL, by Daniel L. Alkon; July, page 70.
MACHINES THAT WALK, by Marc H. Raibert and Ivan E. Sutherland; January, page 44.
MAGNETIC FIELDS IN THE COSMOS, by E. N. Parker; August, page 44.
MAGNETIC FUSION REACTORS, THE ENGINEERING OF, by Robert W. Conn; October, page 60.
MAMMALIAN CELLS, THE LARGE-SCALE CULTIVATION OF, by Joseph Feder and William R. Tolbert; January, page 36.
MEGALITHIC MONUMENTS, THE SOCIAL ARCHAEOLOGY OF, by Colin Renfrew; November, page 152.
MEMORY IN FOOD-HOARDING BIRDS, by Sara J. Shettleworth; March, page 102.
METEORITES, INTERSTELLAR MATTER IN, by Roy S. Lewis and Edward Anders; August, page 66.
MICROBODIES IN THE LIVING CELL, by Christian de Duve; May, page 74.
MICROELECTRONIC PACKAGING, by Albert J. Blodgett, Jr.; July, page 86.
MICROPROGRAMMING, by David A. Patterson; March, page 50.
MITOCHONDRIAL DNA, by Leslie A. Grivell; March, page 78.
MOTTE-AND-BAILEY CASTLE, THE SOCIAL INFLUENCE OF THE, by Michel Bur; May, page 132.
NMR SPECTROSCOPY OF LIVING CELLS, by R. G. Shulman; January, page 86.
NUCLEAR ATTACK, THE UNCERTAINTIES OF A PREEMPTIVE, by Matthew Bunn and Kosta Tsipis; November, page 38.
NUCLEAR-WEAPON-FREE ZONE IN EUROPE, A, by Barry M. Blechman and Mark R. Moore; April, page 37.
OCEAN, THE, by Wallace S. Broecker; September, page 146.
OCEANIC CRUST, THE, by Jean Francheteau; September, page 114.
OPTICAL COMPUTER, THE, by Eitan Abraham, Colin T. Seaton and S. Desmond Smith; February, page 85.
ORGAN PIPES, THE PHYSICS OF, by Neville H. Fletcher and Suszanne Thwaites; January, page 94.
PALM LEAVES, THE DEVELOPMENT OF, by Donald R. Kaplan; July, page 98.
PARTICLES WITH NAKED BEAUTY, by Nariman B. Mistry, Ronald A. Poling and Edward H. Thorndike; July, page 106.
PEKING MAN, by Wu Rukang and Lin Shenglong; June, page 86.
PHYSICAL DISABILITY AND PUBLIC POLICY, by Gerben DeJong and Raymond Lifchez; June, page 40.
PORK PRODUCTION, MODERN, by Wilson G. Pond; May, page 96.
PREY SWITCHING IN A SIMPLE ECOSYSTEM, by Arthur T. Bergerud; December, page 130.
QUANTUM GRAVITY, by Bryce S. DeWitt; December, page 112.
QUARK CONFINEMENT, THE LATTICE THEORY OF, by Claudio Rebbi; February, page 54.
QUARKS AND LEPTONS, THE STRUCTURE OF, by Haim Harari; April, page 56.
RATIONAL COLLECTIVE CHOICE, by Douglas H. Blair and Robert A. Pollak; August, page 88.
RNA, THE PROCESSING OF, by James E. Darnell, Jr.; October, page 90.
SALMONID FISHES AS A NATURAL LIVESTOCK, THE, by Lauren R. Donaldson and Timothy Joyner; July, page 50.
SEED-EATING BEETLE'S ADAPTATIONS TO A POISONOUS SEED, A, by Gerald A. Rosenthal; November, page 164.
SILICON MICROMECHANICAL DEVICES, by James B. Angell, Stephen C. Terry and Phillip W. Barth; April, page 44.
SLIDE FASTENER, THE, by Lewis Weiner; June, page 132.
SLURRY PIPELINES, by Edward J. Wasp; November, page 48.
SMART WEAPONS IN NAVAL WARFARE, by Paul F. Walker; May, page 53.
SOCIAL AMOEBAE, CHEMICAL SIGNALS OF, by John Tyler Bonner; April, page 114.
SOLAR CORONA, THE ACTIVE, by Richard Wolfson; February, page 104.
SPIRAL GALAXIES, DARK MATTER IN, by Vera C. Rubin; June, page 96.
STAVE CHURCHES OF NORWAY, THE, by Petter Aune, Ronald L. Sack and Arne Selberg; August, page 96.
SUDDEN CARDIAC DEATH: A PROBLEM IN TOPOLOGY, by Arthur T. Winfree; May, page 144.
SUPERLATTICES, SOLID-STATE, by Gottfried H. Döhler; November, page 144.
TERMITES, THE CHEMICAL DEFENSES OF, by Glenn D. Prestwich; August, page 78.
TRAUMA, by Donald D. Trunkey; August, page 28.
UNIVERSE, THE FUTURE OF THE, by Duane A. Dicus, John R. Letaw, Doris C. Teplitz and Vigdor L. Teplitz; March, page 90.
UNIVERSE, THE LARGE-SCALE STRUCTURE OF THE, by Joseph Silk, Alexander S. Szalay and Yakov B. Zel'dovich; October, page 72.
URBANIZATION IN THE U.S., THE SLOWING OF, by Larry Long and Diana DeAre; July, page 33.
VACCINES, SYNTHETIC, by Richard A. Lerner; February, page 66.
VECTOR FOR INTRODUCING NEW GENES INTO PLANTS, A, by Mary-Dell Chilton; June, page 50.
VISUAL ILLUSIONS, THE INTERPRETATION OF, by Donald D. Hoffman; December, page 154.
VISUAL PROCESSES, HIDDEN, by Jeremy M. Wolfe; February, page 94.
VOLCANIC CALDERAS, GIANT, by Peter Francis; June, page 60.
WHALES AND PORPOISES, THE PHYSIOLOGICAL ECOLOGY OF, by John W. Kanwisher and Sam H. Ridgway; June, page 110.
X-RAY BACKGROUND, THE ORIGIN OF THE COSMIC, by Bruce Margon; January, page 104.

METAMAGICAL THEMAS

Cooperation evolves. Computer tournaments of the Prisoner's Dilemma suggest how; May, page 16.
Cooperation is tested through a lottery, The calculus of; June, page 14.
Lisp concludes with a gargantuan Italian feast, In which a discourse on the language; April, page 14.
Lisp: the chosen language of artificial in-

telligence, The pleasures of; February, page 14.
Lisp, the language of artificial intelligence, Tripping the light recursive in; March, page 22.
Parquet deformations: patterns of tiles that shift gradually in one dimension; July, page 14.
Self-replicating structures, Virus-like sentences and; January, page 14.

MATHEMATICAL GAMES

Knots, plus the results of Douglas Hofstadter's Luring Lottery, The topology of; September, page 18.
Tasks you cannot help finishing no matter how hard you try to block finishing them; August, page 12.

COMPUTER RECREATIONS

Computation, Introducing a department concerned with the pleasures of; October, page 22.
Finite-state machine, a minimal model of mousetraps, ribosomes and the human soul, On the; December, page 19.
Literature into drivel, A progress report on the fine art of turning; November, page 18.

THE AMATEUR SCIENTIST

Billiards and pool), The physics of the follow, the draw and the masse (in; July, page 124.
Caustics: mathematical curves generated by light shined through rippled plastic; September, page 190.
Coffee is brewing in the ibrik, How lenses can be made out of ice, and what happens when; April, page 132.
Ice, and what happens when coffee is brewing in the ibrik, How lenses can be made out of; April, page 132.
Knot will hold or slip, In which simple equations show whether a; August, page 120.
Oil or other substances are placed on water, Funny things happen when drops of; December, page 164.
Physics, Simple and vivid demonstrations of advanced concepts in; January, page 130.
Physics while scared to death (on a falling roller coaster), Thinking about; October, page 162.
Polarizing filters? What causes the color in plastic objects stressed between two; June, page 146.
Thermal oscillators: systems that seesaw, buzz or howl under the influence of heat; February, page 146.
Traffic-light system from the outside looking in, How to analyze a city; March, page 138.
Water Striders, the insects that walk (and run) on water, Looking into the ways of; November, page 188.
Wine? What causes the "tears" that form on the inside of a glass of; May, page 162.

Friends of Libraries U.S.A.

We're a new national organization dedicated to encouraging and assisting the formation and development of Friends of Libraries groups throughout the country.

Won't you join us?

Name

Address

City State Zip

Library

Address

City State Zip

For an application form, write:
Friends, ALA
50 E. Huron
Chicago, Illinois 60611

SAVE 40% ON

CASIO
FX 700
computer system

NEW!



price includes:
computer
cassette interface
printer
cassette program

125.00

Mfr. sugg. ret. 209.85

EASY TO LEARN, easy to use basic language computer has built-in, non-volatile, variable memory (maximum 1568 program steps and max. 222 memories). Up to 10 program areas. Large variety of cassette-based software available now.

One-key commands permit quick BASIC input — 54 keyboard based on ASCII standard arrangement.

CALL TOLL FREE 800-621-1269
EXCEPT Illinois, Alaska, Hawaii

Corp. Accts. Invited. Min. Ord. $15.00 Mastercard or Visa by mail or phone. Mail Cashier's Check, Mon. Ord. Pers. Check (2wks to clr.) Add $4.00 1st item. **(AK, HI, P.R., Canada add $10.00 first item)** $1.00 ea. add'l shpg. & handl. Shpts. to IL address add 6% tax. Prices subj. to change. **WRITE for free catalog. 30 days return policy for defective replacements only. ALL ELEK-TEK MDSE. IS BRAND NEW, 1ST QUALITY AND COMPLETE.**

ELEK-TEK, inc.
6557 N. Lincoln Ave., Chicago IL 60645
(800) 621-1269 (312) 677-7660

INVEST YOURSELF



VITA Putting Resources to Work for People

1815 North Lynn Street, Arlington, Virginia 22209-2079, USA

A windmill to pump water for "salt farming" in India. More efficient woodburning stoves for the Sahel. Photovoltaic irrigation pumps for the Somali refugee camps.

All these are solutions to technical problems in developing countries. Devising such solutions is no simple task. To apply the most advanced results of modern science to the problems of developing areas in a form that can be adopted by the people requires the skills of the best scientists, engineers, farmers, businessmen—people whose jobs may involve creating solid state systems or farming 1000 acres, but who can also design a solar still appropriate to Mauritania or an acacia-fueled methane digester for Nicaragua.

Such are the professionals who volunteer their spare time to Volunteers in Technical Assistance (VITA), a 20 year old private, non-profit organization dedicated to helping solve development problems for people world-wide.

Four thousand VITA Volunteers from 82 countries donate their expertise and time to respond to the over 2500 inquiries received annually. Volunteers also review technical documents, assist in writing VITA's publications and bulletins, serve on technical panels, and undertake short-term consultancies.

Past volunteer responses have resulted in new designs for solar hot water heaters and grain dryers, low-cost housing, the windmill shown above and many others. Join us in the challenge of developing even more innovative technologies for the future.

BIBLIOGRAPHY

Readers interested in further explanation of the subjects covered by the articles in this issue may find the following lists of publications helpful.

COMPUTER RECREATIONS

COMPUTATION: FINITE AND INFINITE MACHINES. Marvin L. Minsky. Prentice-Hall, Inc., 1967.

MATHEMATICAL STRUCTURES FOR COMPUTER SCIENCE. Judith L. Gersting. W. H. Freeman and Company, 1967.

ALGORITHMS + DATA STRUCTURES = PROGRAMS. Niklaus Wirth. Prentice-Hall, Inc., 1976.

INTRODUCTION TO AUTOMATA THEORY, LANGUAGES, AND COMPUTATION. John E. Hopcroft and Jeffrey D. Ullman. Addison-Wesley Publishing Company, 1979.

MODERN ICEBREAKERS

THE NORTHEAST PASSAGE AS A COMMERCIAL WATERWAY. Terence Armstrong in *Vega 1880–Ymer 1980: Årsbok 1980 Svenska Sällskapet för Antropologi och Geografi,* pages 86–129; 1980.

SHIPPING ROUTES, ICE COVER AND YEAR-ROUND NAVIGATION IN THE CANADIAN ARCTIC. Balaram B. Dey in *Polar Record,* Vol. 20, No. 129, pages 549–559; September, 1981.

THE SOVIET NORTH IN THE ELEVENTH FIVE-YEAR PLAN, 1981–85. Terence Armstrong in *Polar Record,* Vol. 21, No. 130, pages 66–69; January, 1982.

IO

MELTING OF IO BY TIDAL DISSIPATION. S. J. Peale, P. Cassen and R. T. Reynolds in *Science,* Vol. 203, No. 4383, pages 892–894; March 2, 1979.

THE ROLE OF SO_2 IN VOLCANISM ON IO. B. A. Smith, E. M. Shoemaker, S. W. Kieffer and A. F. Cook II in *Nature,* Vol. 280, No. 5725, pages 738–743; August 30, 1979.

THE GALILEAN SATELLITES AND JUPITER: VOYAGER 2 IMAGING SCIENCE RESULTS. Bradford A. Smith, Laurence A. Soderblom et al. in *Science,* Vol. 206, No. 4421, pages 927–950; November 23, 1979.

THE GALILEAN MOONS OF JUPITER. Laurence A. Soderblom in *Scientific American,* Vol. 242, No. 1, pages 88–100; January, 1980.

HEAT FLOW FROM IO (JI). Dennis L. Matson, Gary A. Ransford and Torrence V. Johnson in *Journal of Geophysical Research,* Vol. 86, No. B3, pages 1664–1672; March 10, 1981.

SATELLITES OF JUPITER. Edited by D. Morrison. University of Arizona Press, 1982.

TWO CLASSES OF VOLCANIC PLUMES ON IO. Alfred S. McEwen and Laurence A. Soderblom in *Icarus,* Vol. 55, No. 2, pages 191–217; August, 1983.

AN EARLY IRON AGE FARM COMMUNITY IN CENTRAL EUROPE

EUROPEAN PREHISTORY. Sarunas Milisauskas. Academic Press, 1978.

FROM BRONZE AGE TO IRON AGE: ECONOMIC, HISTORICAL AND SOCIAL CONSIDERATIONS. R. Peroni in *Italy before the Romans: The Iron Age, Orientalizing, and Etruscan Periods,* edited by David and Francesca R. Ridgway. Academic Press, 1979.

RURAL ECONOMY IN THE EARLY IRON AGE: EXCAVATIONS AT HASCHERKELLER 1978–1981. Peter S. Wells. Peabody Museum of Harvard University, 1983.

THE DNA HELIX AND HOW IT IS READ

DNA: RIGHT-HANDED AND LEFT-HANDED HELICAL CONFORMATIONS. Alexander Rich, Gary J. Quigley and Andrew H.-J. Wang in *Biomolecular Stereodynamics: Vol. 1,* edited by R. H. Sarma. Adenine Press, 1981.

THE ANATOMY OF A-, B-, AND Z-DNA. Richard E. Dickerson, Horace R. Drew, Benjamin N. Conner, Richard M. Wing, Albert V. Fratini and Mary L. Kopka in *Science,* Vol. 216, No. 4545, pages 475–485; April 30, 1982.

MECHANICS OF SEQUENCE-DEPENDENT STACKING OF BASES IN *B*-DNA. C. R. Calladine in *Journal of Molecular Biology,* Vol. 161, No. 2, pages 343–352; October 25, 1982.

ORDERED WATER STRUCTURE AROUND A *B*-DNA DODECAMER: A QUANTITATIVE STUDY. Mary L. Kopka, Albert V. Fratini, Horace R. Drew and Richard E. Dickerson in *Journal of Molecular Biology,* Vol. 163, No. 1, pages 129–146; January 5, 1983.

SEQUENCE-DEPENDENT CONFORMATION OF AN *A*-DNA DOUBLE HELIX. Z. Shakked, D. Rabinovich, O. Kennard, W. B. T. Cruse, S. A. Salisbury and M. A. Viswamitra in *Journal of Molecular Biology,* Vol. 166, No. 2, pages 183–201; May 15, 1983.

BASE SEQUENCE AND HELIX STRUCTURE VARIATION IN *B* AND *A* DNA. Richard E. Dickerson in *Journal of Molecular Biology,* Vol. 166, No. 3, pages 419–441; May 25, 1983.

QUANTUM GRAVITY

GENERAL RELATIVITY: AN EINSTEIN CENTENARY SURVEY. Edited by S. W. Hawking and W. Israel. Cambridge University Press, 1979.

QUANTUM GRAVITY II. Edited by C. J. Isham, R. Penrose and D. W. Sciama. Oxford University Press, 1981.

PREY SWITCHING IN A SIMPLE ECOSYSTEM

THE DISTRIBUTION AND ABUNDANCE OF ANIMALS. H. G. Andrewartha and L. C. Birch. University of Chicago Press, 1954.

ECOLOGICAL GENETICS. E. B. Ford. John Wiley & Sons, Inc., 1965.

OF PREDATION AND LIFE. Paul L. Errington. Iowa State University Press, 1967.

THE AERODYNAMICS OF HUMAN-POWERED LAND VEHICLES

MAN-POWERED VEHICLE DESIGN CRITERIA. Chester R. Kyle and Walter Edelman in *Proceedings of the Third International Conference on Vehicle System Dynamics held at Virginia Polytechnic Institute and State University, August 12–15, 1974, Blacksburg, Virginia,* edited by H. K. Sachs. Swets & Zeitlinger, Amsterdam, 1975.

PROCEEDINGS OF THE FIRST HUMAN POWERED VEHICLE SCIENTIFIC SYMPOSIUM. Edited by Alan Abbott. IHPVA, P.O. Box 2068, Seal Beach, Calif. 90740, 1981.

BICYCLES & TRICYCLES: AN ELEMENTARY TREATISE ON THEIR DESIGN AND CONSTRUCTION. Archibald Sharp. Reprint of 1896 edition by the MIT Press, 1982.

BICYCLING SCIENCE. Frank R. Whitt and David G. Wilson. The MIT Press, 1982.

THE INTERPRETATION OF VISUAL ILLUSIONS

DIFFERENTIAL GEOMETRY OF CURVES AND SURFACES. Manfredo P. do Carmo. Prentice-Hall, Inc., 1976.

VISION: A COMPUTATIONAL INVESTIGATION INTO THE HUMAN REPRESENTATION AND PROCESSING OF VISUAL INFORMATION. David Marr. W. H. Freeman and Company, 1982.

THE MODULARITY OF MIND. Jerry Fodor. The MIT Press, 1983.

THE AMATEUR SCIENTIST

THE HYDROPHOBIC EFFECT: FORMATION OF MICELLES AND BIOLOGICAL MEMBRANES. Charles Tanford. John Wiley & Sons, Inc., 1980.

INTERATOMIC AND INTERMOLECULAR FORCES. John P. McTague in *Encyclopedia of Physics,* edited by Rita G. Lerner and George L. Trigg. Addison-Wesley Publishing Company, Inc., 1981.

MARYGROVE
LIBRARY
Detroit, Mi.